*Tesis Doctoral*

# SEEING HATE FROM AFAR

## The Concept of Affective Polarization Reassessed

Tesis presentada por

Manuel Almagro Holgado

para optar al grado de Doctor con Mención Internacional
en el programa de Doctorado en Filosofía (BO2.56.1)

Director: Neftalí Villanueva Fernández
Tutor: Manuel de Pinedo García

UNIVERSIDAD
DE GRANADA

Facultad de Filosofía y Letras
Departamento de Filosofía I

Editor: Universidad de Granada. Tesis Doctorales
Autor: Manuel Almagro Holgado
ISBN: 978-84-1117-014-7
URI: http://hdl.handle.net/10481/70432

# Contents

# Agradecimientos
# Acknowledgments

Esta tesis doctoral es en mayor o menor medida de todas y cada una de las personas que conforman el grupo de investigación al que estoy adscrito y al que admiro enormemente, con excepción de los errores y los puntos flacos de la misma; de ellos solo yo soy responsable.

Hay muchísimas personas a las que quiero dar las gracias públicamente por haber contribuido de una manera u otra a que este trabajo haya tenido no solo un principio sino también un final. Ha sido todo un privilegio tener la oportunidad de realizar una tesis doctoral, especialmente en las condiciones en las que esta tesis se ha llevado a cabo, y por ello estoy infinitamente agradecido. Me gustaría empezar dando las gracias a quienes han posibilitado que este trabajo tenga un principio y terminar con las personas que también han estado en su etapa final.

Mi madre nació en 1943 y mi padre en 1941. Ambos lidiaron con la ruina de la posguerra. Desde los cinco años, mi padre tuvo que trabajar en diferentes casas exclusivamente a cambio de comida y nunca tuvo la oportunidad de ir a la escuela. Mi madre tuvo un poco de mejor fortuna y pudo recibir la formación elemental, aunque pronto tuvo que trabajar y abandonar sus estudios. Mi madre y mi padre han hecho todo cuanto ha estado en sus manos para que mi hermana, mi hermano y yo tuviésemos las oportunidades que ellos no tuvieron. Esta tesis es una prueba de su triunfo.

Mi hermana y mi hermano, ocho y diez años mayores que yo, son en buena medida responsables de mi camino. La excepcionalidad de mi hermana abrió una puerta hasta entonces bloqueada para mi familia y para prácticamente todo nuestro entorno. A finales de los 90 se marchó a estudiar a la Universidad de Granada gracias a la insistencia y ayuda de quienes atestiguaron su excepcionalidad. Mi hermano me animó a que, a los casi dieciocho años, retomara los estudios que había abandonado a los quince, y siempre estuvo a mi lado en los momentos más difíciles de mi adolescencia. Sin mi hermana y mi hermano esta tesis nunca habría llegado a producirse.

Mi amigo Fares ha sido un ejemplo de fortaleza, superación, humor y positividad. Soy extremadamente afortunado de que hayamos sido inseparables desde que tengo memoria. Su familia me acogió y cuidó como a un hijo más, y a ellos les debo muchísimo. A mi barrio, el polígono de Ceuta, le estoy enormemente agradecido, no solo por las magníficas personas con las que he tenido la suerte de criarme sino también por todas las experiencias, de diverso tipo, que he vivido allí y que me han enseñado tanto. Estoy convencido de que hay cachitos de todo ello en este trabajo.

A la asociación de Ceuta *Casa de Estudios CE-70* le estoy inmensamente agradecido por haber proporcionado buena parte de los recursos materiales sin los que no podría haber cursado mis estudios de grado y de máster en Granada. En especial le doy las gracias a todas mis paisanas con las que compartí experiencias en esta asociación, tanto en los veranos de trabajo en la feria de Ceuta como en la convivencia y la diversión en Granada.

Formo parte del grupo de personas que afirman que la música, y particularmente el rap, les salvó la vida. La música ha sido un motor constante en mi vida y, cómo no, también lo ha sido en el desarrollo de este trabajo. No podía faltar una mención a ello aquí. En especial, me gustaría dar las gracias a mi paisano Soto Asa por su música, que ha sido la banda sonora de mis últimos años de trabajo en esta tesis y me ha mantenido conectado a mis orígenes. La primera parte del título de esta tesis es un guiño a uno de los versos de su canción “P3n4”. Arriba la casa!

Durante los años del grado y del máster conocí a muchas personas que hicieron de Granada un lugar inmejorable para mí. Mil gracias a mis amigas Candela Caballero, Elena Díaz, Cecilia Prieto, Javier Ramírez, Ismael Romero, Fernando Sánchez y Juanmi de la Torre por tantas barbacoas y tantos días y noches de fiesta, conciertos, risas y conversaciones.

Durante el máster interuniversitario en lógica y filosofía de la ciencia conocí a unas cuantas personas que han dejado huella en mí. Ricardo Grande, además de compañero de fiesta, fue uno de los compañeros de máster con quien más he discutido sobre filosofía. Daniel Galdeano, desde que lo conocí allá por el 2016, también en el máster, se ha convertido en una de las personas a las que más admiro y aprecio. Con él compartí mi primera presentación en un congreso mientras aprendíamos a *ver-como* el camino. Le doy las gracias a ellos y a todas las personas, alumnado y profesorado, que coincidimos ese año en Salamanca. Fue un año increíble.

Estoy muy agradecido al Programa Operativo de Empleo Juvenil (POEJ) 2014-2020, desarrollado en el marco del Fondo Social Europeo (FSE), que me permitió disfrutar de un contrato de personal técnico de apoyo a la investigación y a la gestión de la I+D+i entre 2017 y 2018 para trabajar en diversos proyectos con Juan Antonio Nicolás y Ana Ramírez.

Desde 2018 tengo el privilegio de disfrutar de un contrato de Formación del Personal Investigador (FPI: BES-2017-079933), asociado al proyecto "Expresivismos Contemporáneos y la Indispensabilidad del Vocabulario Normativo: Alcance y Límites de la Hipótesis Expresivista" (FFI2016-80088-P), financiado por el Ministerio de Economía, Industria y Competitividad, que me ha permitido dedicarme a tiempo completo a desarrollar esta investigación y participar en congresos nacionales e internacionales. Estoy enormemente agradecido por ello, ha sido un sueño cumplido.

He tenido también la enorme suerte de impartir docencia en la Universidad de Granada y beneficiarme de tal experiencia. Le doy las gracias al alumnado de las asignaturas *Implicaciones Sociales de la Biotecnología* del Grado en Biotecnología (2019-2020), *Filosofía* del Grado en Historia (2019-2020) y *Filosofía de la Mente* del Grado en Filosofía (2020-2021). Me habéis enseñado muchísimo. En especial, me gustaría dar las gracias a las magníficas alumnas Clara Cámara y Alicia Ibáñez por haberme permitido colaborar en la dirección de sus trabajos de fin de grado y fin de máster respectivamente. Ha sido todo un lujo para mí.

Durante estos años he tenido la fortuna de encontrarme con gente maravillosa de la que he aprendido muchísimo y con la que he compartido experiencias inolvidables. Algunas de ellas son Gloria Andrada, Bethany Ansell, Virginia Ballesteros, Moisés Barba, Cristina Bernabeu, Bruno Borge, Cristina Borgoni, Marta Cabrera, Carla Carmona, Cristina Corredor, Marga Díaz, Jordi Fairhust, Hadis Farokhi, Ainhoa Fernández, Antonio Gaitán, Laura Giraldo, Javier González, Marta Jorba, Andrés Luna, Jesús Navarro, Marina Ortega, Saúl Pérez, Claudia Picazo, Daniel Pino, Jon Rueda, Mariano Sanjuán, Leonie Smith, Jesús Vega, Hugo Viciana y José Manuel Viejo. Agradezco a Lilian Bermejo, Fernando Martínez y José Luis Moreno sus observaciones y sugerencias en alguna de las etapas iniciales de este trabajo como parte de las diferentes comisiones de evaluación académica en las que he tenido la fortuna de estar. Gracias también a las diversas audiencias a las que he tenido la suerte de presentar partes de esta investigación por sus comentarios, observaciones y recomendaciones. Especial agradecimiento a Rodrigo Díaz, Ivar Hannikainen y Jimena Zapata, con quienes las conversaciones se han convertido además en trabajo académico. Ivar ha sido una fuente constante de aprendizaje y diversión durante los últimos años. Muchas gracias por todo.

Entre enero y abril de 2021 he tenido el privilegio de hacer una estancia de investigación en la Universidad Ludwig-Maximilian (LMU) de Múnich, bajo la supervisión de Ophelia Deroy. Le estoy muy agradecido a Ophelia y a todas las personas del laboratorio interdisciplinar *Cognition, Values, Behaviour* (CVBE) por haberme acogido y permitido presentar y discutir mi trabajo. A pesar de las circunstancias pude aprender muchísimo de todas ellas. Gracias especialmente a Jimena por haber facilitado mi llegada y mi estancia en Múnich y, sobre todo, por haberme permitido seguir y discutir su interesantísimo trabajo.

No puedo estar más agradecido al grupo al que tengo la suerte de pertenecer, también conocido como Granada Gang. La calidad humana, intelectual y profesional tan extraordinaria de la gente que lo compone, la forma de entender la filosofía y la orientación práctica y política de buena parte de sus trabajos han hecho del grupo un lugar inmejorable, una fuente inagotable de crecimiento académico y personal. En primer lugar, quiero dar las gracias a Juan José Acero, María José Frápolli y Esther Romero por su incalculable contribución a mi desarrollo desde los inicios de mi formación académica. Millones de gracias a Alba, Amalia, Andrés,

Dani, David, Edu, Francesco, Javier “Xavi”, José, José Ramón “Tori”, Liñán, Llanos, Manolo “Manoloheras”, Miguel “Trunkel”, Mirco, Nemesio, Palma, Pedro y Víctor por haber generado las condiciones idóneas en las que comenzar y desarrollar esta investigación. Doy las gracias también a la parte extraoficial del grupo, Alberto, Cristina, Layla, Lorena, Mar, María, María José, Rocío y Sara, por su influencia y por los momentos compartidos. David, Manoloheras y Tori han sido una guía y una influencia muy grande para mí durante estos años, y Víctor ha sido prácticamente un hermano. Muchas gracias por estar siempre ahí y por enseñarme tanto. Alba, Amalia, Edu, Llanos, Trunkel y Xavi han sido las compañeras y amigas con las que más unido he estado durante el doctorado. Muchísimas gracias por los viajes, congresos, fiestas, barbacoas, discusiones y risas que han llenado estos años y que le han dado tanto sentido. Mención especial a Llanos y Trunkel por haber discutido tanto de filosofía conmigo y por haber aportado siempre una perspectiva fina, y a Xavi por, entre otras muchas cosas, haberme echado una mano con la portada de la versión libro de esta tesis.

Mis directores Manuel de Pinedo y Neftalí Villanueva han sido y son mi referencia no solo de lo que es ser filósofo, profesor e investigador, sino también de lo que es ser jefe, compañero, amigo y persona. Durante todos estos años mi admiración por ellos no ha hecho más que crecer. Ellos han confiado en mí, me han guiado, me han enseñado, me han inspirado, me han aconsejado, me han cuidado y han sacado lo mejor de mí, siempre con amabilidad y con libertad absoluta para que yo eligiera mi propio camino. Estoy convencido de que han hecho por mí y por el grupo mucho más de lo que soy capaz de reconocer. Gracias infinitas por todo lo que hacéis, por haberme dado la oportunidad de llevar a cabo este proyecto y por haber sido no solo mis directores sino también mis amigos. Me he esforzado al máximo para tratar de compensar vuestro compromiso, aunque eso nunca será posible.

A Encarnación y Julia les doy las gracias por haberme acogido como a uno más de la familia todos estos años y por haberse preocupado tanto por mí. Sois estupendas. Finalmente quiero dar las gracias a mi pareja, Ana, con quien tengo la enorme fortuna de compartir una vida plena de comprensión, complicidad, cuidado y crecimiento. Siempre me has apoyado en todas mis decisiones, me has

animado a que me esfuerce y me has comprendido y consolado en mis malas rachas. La capacidad, honestidad, integridad, bondad y compromiso que expresan tus acciones hacen que mejore diariamente. Tu alegría es la luz de mi vida. Sin ti no habría siquiera empezado este camino. Gracias por tanto. Solo me siento en casa cuando estoy contigo.

# Publications

Almagro, Manuel, Javier Osorio and Neftalí Villanueva (Forthcoming). Weaponized Testimonial Injustice. *Las Torres de Lucca.*

Almagro, Manuel, Ivar R. Hannikainen and Neftalí Villanueva (Forthcoming). Whose Words Hurt? Contextual Determinants of Offensive Speech. *Personality and Social Psychology Bulletin.*

Almagro, Manuel and Alba Moreno (Forthcoming). Affective Polarization and Testimonial and Discursive Injustice. In Bordonaba, David, Víctor Fernández & José R. Torices (eds.), *The Political Turn in Analytic Philosophy*, De Gruyter.

Almagro, Manuel, Llanos Navarro-Laespada and Manuel de Pinedo (Forthcoming). Is Testimonial Injustice epistemic? Let me count the ways. *Hypatia. A Journal of Feminist Philosophy.*

Almagro, Manuel and Neftalí Villanueva (2021). Qué decir y qué esperar cuando hablamos de la pandemia. *Revista de la Sociedad de Lógica, Metodología y Filosofía de la Ciencia en España*, Número especial: Filosofía en tiempos de pandemia, pp. 64-71.

Almagro, Manuel and Neftalí Villanueva (2021). Polarización y tecnologías de la información: radicales Vs. extremistas. *Dilemata. Revista Internacional de Éticas Aplicadas*, 34: 51-69.

Almagro, Manuel (2020). Límites de la noción de affordance y de la concepción de lo mental en el marco de la psicología ecológica. *Teorema, Revista Internacional de Filosofía*, 39(1): 135-149.

Almagro, Manuel and Víctor Fernández Castro (2020). The Social Cover View: a non-epistemic approach to mindreading. *Philosophia*, 48(2): 483-505.

Almagro, Manuel and Neftalí Villanueva (2020). Por qué debe importarnos la polarización afectiva. *The Conversation*, ISSN: 2201-5639.

Almagro, Manuel and Neftalí Villanueva (2020). Desinformados y ofendidos. *El Salto*.

Almagro, Manuel (2019). La polarización política: polarización expresiva o en actitudes. In *Revista de la SLMFCE.* Número Extraordinario Congreso de Posgrado 2019, pp. 41-44. ISSN: 2695-480X

Almagro, Manuel and Rodrigo Díaz (2019). You are just being emotional! Testimonial injustice and folk-psychological attributions. *Synthese*, Special Issue "Folk psychology: Pluralistic approaches". DOI:10.1007/s11229-019-02429-w

Almagro, Manuel (2019). Book review of J. J. Acero (ed.), Guía Comares de Wittgenstein. *Teorema, Revista Internacional de Filosofía*, 38(2): 137-142.

Almagro, Manuel (2019). Affordances e Injusticia Social. *Ciencia Cognitiva*, 13:2, 35-37.

Almagro, Manuel (2018). El lenguaje inclusivo frente a la RAE. In *CTXT Contexto y Acción.* Número 176.

Almagro, Manuel and José Ramón Torices (2018). The Nature of (Covert) Dogwhistles, in. *IX Conference of the Spanish Society for Logic, Methodology and Philosophy of Science*, 50(148).

Almagro, Manuel (2018). Verdad sobre la no existencia: un problema para la teoría reduccionista de Tim Crane. Nota crítica sobre The Objects of Thought. *Crítica, revista hispanoamericana de filosofía*, 50(148): 99-113.

Almagro, Manuel (2017). Wittgenstein y la hipótesis del espectro visual invertido. *Análisis, revista de investigación filosófica*, 4(1): 77-92.

Almagro, Manuel (2017). El expresivismo clásico y los límites de la interpretación naturalista de la intencionalidad en el programa de Grice. *Areté, Revista de Filosofía*, 29(1): 7-27.

# Summary

According to the online search engine Google Ngram Viewer, the items related to the terms 'political polarization' and 'political polarisation' have enormously increased since 2010. The topic of political polarization has received a significant amount of public and academic attention. Certainly, some authors have focused on the benefits of the rise of political polarization (Stavrakakis 2018), but must of them have focused on the negative consequences of the rise polarization for the well-functioning of democratic institutions (Carothers & O'Donohue 2019; McCoy & Somer 2019). As Levitsky and Ziblatt point out, political polarization can kill democracy (Levitsky & Ziblatt 2018), and it can be merely the result of our tendency to overdo democracy and politics (Talisse 2019), which makes us more radicalized (Sunstein 2017) and behave like arrogant know-it-alls (Lynch 2019). But regardless of the diagnosis offered about the current situation of many democratic societies, it seems important to clarify exactly what we talk about when we talk about political polarization.

The term 'political polarization' can be used to refer to many different phenomena. In addition to the mechanisms that are related to the rise of polarization and the consequences of it, the label is often used to refer to different forms and types of polarization. Two prominent notions are the concepts of ideological polarization and affective polarization. The former is commonly conceived as the separation of the ideological beliefs in a population, either in terms of the distance of the political opinions in an ideological distribution, the dispersion of them, or in terms of any other parameter used to represent the beliefs of a population in an ideological spectrum (Bramson et al. 2017). Affective polarization, on the other

hand, is commonly understood in terms of positive feelings toward the in-group and negative feelings toward the out-group (Iyengar et al. 2012).

Although ideological polarization seems to differentiate from affective polarization in the type of mental state that it deals with, both concepts target polarization by employing self-report questionnaires. In order to know what people believe and feel, the standard tools used to measure polarization directly ask participants to indicate their own mental states. Based on this, it seems that there are at least two philosophical assumptions behind both concepts. First, both concepts assume the difference between the nature of belief-like mental states and desire-like states, which explains the difference between both types of polarization. Second, both concepts endorse the first-person authority thesis, according to which the speaker's sincerity guarantees the truth of her mental self-ascription, and that's the reason why, in order to know what people believe and feel, respondents are directly asked about their own mental states.

A first concern related to ideological polarization is that its relation with the pernicious consequences of the rise of polarization is not obvious. Diversity of opinion seems to be one of the essential pillars of any democracy, and without further explanation it is not at all obvious why a society whose political beliefs are clustered into blocks or dispersed along an ideological distribution will be a democracy in danger. A second concern related to this notion is that it seems to sublimate the middle point of an ideological distribution, but it is not obvious why the beliefs located in the middle of a distribution are necessarily preferable to the ones located at one extreme. Increased identification with a "centrist" ideology could be a problem for democracy as well if, for example, such people were systematically unwilling to coordinate and to live together with people with different political ideas. In this sense, ideological polarization does not seem to account for the problems associated with the rise of political polarization. Affective polarization, on the other hand, does manage to explain why increased polarization is a problem for the proper functioning of democracy: the increase in positive feelings such as sympathy toward the in-group and negative feelings such as hostility toward the out-group explains the difficulty in consensus and coordination with "the others".

Although affective polarization seems to be better positioned than ideological polarization to account for the dangers associated with the rise of political polarization, this notion also exhibits certain explanatory limitations. First, the diagnosis that follows from affective polarization, as it is commonly understood in the literature, is that political polarization is not about what we believe, but about what we feel. This explanation hardly fits with the tendency to produce rational discourse, to offer arguments that support the ideological beliefs of the group with which one identifies, exhibited by a polarized population. People do not simply hate their opponents, but think that evidence is on their side and that the others are clearly wrong. Second, this notion of polarization does not seem to be able to account for certain polarization processes in which the feelings of a population remain stable and yet the level of imperviousness to arguments coming from the opposing sides increases.

Beyond these difficulties, the philosophical assumptions behind both notions of polarization seem to be highly challengeable. A good number of empirical studies (and everyday experiences), together with our intuitions in many cases, suggest that we frequently fail at identifying our own mental state, be it a belief or a feeling, even when we self-ascribe it in a sincere manner (Schwitzgebel 2008, 2011a,b). And it is also far from clear that beliefs and feelings are radically different mental states with respect to their link to action. The first of these two objections is particularly critical, because it suggests that nothing ensures that polarization measurement tools accurately measure what they target.

The aim of this dissertation is to offer a notion of polarization free of those philosophical assumptions. In particular, we seek a notion that avoids the difficulties pointed out before and that can account for some polarization processes that go unnoticed to the notions of ideological polarization and affective polarization, as they are commonly understood. For this purpose, we propose five desiderata that, in our opinion, an adequate notion of polarization must meet. First, this notion must be able to explain the pernicious effects of the rise of polarization for democracy. Second, it must be consistent with our best evidence. Third, it must not explain the rise of polarization in terms of irrationality. Fourth, it must accommodate the distinction between saying that someone believes or feels that p

and the feelings and beliefs that someone actually has. Finally, it must allow for intervention.

This dissertation is structured in eight chapters and has two main arguments, one negative and one positive. The negative argument holds that the concepts of ideological polarization and affective polarization, in the way they are commonly understood, have an array of limitations that can hardly be avoided, and that both concepts don't actually measure the mental states that they try to measure. The positive argument holds that the tools employed to measure affective polarization actually target not (only) people's feelings, but the attitudes people *express*, more specifically those connected with their level of credence in the core beliefs of the political group that they identify with. Understood this way, affective polarization meets the desiderata and avoids the problems associated with the notion.

To reach our objective, we start by making a state of the art of political polarization of beliefs (chapter 2) by drawing conceptual distinctions around three categories: general forms of polarization, types of polarization, and conceptions of polarization. A crucial distinction here is the distinction between belief content and degree of belief (Talisse 2019), i.e., the distinction between what the population believes and how it believes it. This distinction is at the basis of two quite distinct conceptions of polarization of beliefs. Ideological polarization focuses exclusively on what people believe, i.e., belief content, represented by one of the meanings given by the dictionary of the *Real Academia Española* (RAE) for the verb ‘to polarize: “to orient in two opposite directions”. However, there is another way of understanding polarization of beliefs, based not on what people believe, but on how people believe what they believe, represented by another of the meanings for the same verb: “to concentrate attention or mood on something”.

We then introduce the notion of affective polarization as it is usually understood in the literature, present and discuss both the limitations that the notions of ideological polarization and affective polarization face and the philosophical assumptions they share, and introduce a set of conditions that a suitable notion of polarization must meet (chapter 3).

To satisfy the requirement of evidence on how we polarize, we review the relevant literature about the mechanisms, of a different nature, related to the rise of polarization. Crucially, we analyze the compatibility between these mechanisms and polarization of beliefs, understood both in terms of belief content and in terms of degree of belief (chapter 4). Ideological polarization, which has to do with belief contents, seems unable to accommodate the evidence, or at least seems to be worse positioned to do so than polarization of beliefs understood as shifts in the degree of belief. This evidence, moreover, seems more compatible with the idea that polarization is a rational process than with the irrational story of polarization (Dorst 2020).

To satisfy the requirement of disanalogy between what we say that we believe and feel and what we actually believe and feel, we need a theory about dispositional mental state ascriptions that can accommodate the difference between self-attributing a mental state and expressing a mental state, and that rejects the first-person authority thesis without violating our intuitions in a number of natural cases of belief self-ascriptions. In chapter 5 we discuss the family of theories committed to the idea that the function of our mental self-ascriptions is to describe a state of affairs, and argue that these theories cannot account for the disanalogy.

In chapter 6 we offer an approach to mental attributions, based on our interpretation of a number of Wittgenstein's remarks, that satisfies the requirement of disanalogy. According to this approach, mental attributions do not accomplish a descriptive function, and their truth-values do not depend on whether a particular state of affairs, internal or external to the individual, is the case, but on the compatibility of their truth with the relevant features of the context. Thus, being in a state of mind is not a matter of fact, but a normative issue: when we ascribe a belief to ourselves or to other people, we acquire a set of conceptual commitments linked to certain courses of action, and its truth will depend on its compatibility with the salient features of the context. This approach accommodates the possibility of error in ascribing a mental state.

In chapter 7 we draw on some contemporary expressivisms to argue that affective polarization is better positioned than ideological polarization to measure the relevant attitudes of a population, i.e., our conceptual commitments especially linked to action. The main argument for that is that the tools usually employed to measure affective polarization involve evaluative language, and through the evaluative use of language we express our commitments, our attitudes, conceptually linked to our world-picture. That is, through our evaluations we commonly express our affective attitudes, which sometimes do not fit with our self-ascribed mental states. Crucially, these attitudes, especially linked to action, are closely connected with our degree of belief: we express the level of credibility we have in certain core beliefs of the ideological group we identify with. This move allows us to reassess the concept of affective polarization, which we call *polarization in attitudes* to distinguish it from the traditional way in which the concept is understood, in a way that avoids the problems associated with the concept of affective polarization. Besides, our concept don't entail the first-person authority thesis, meets the desiderata and can account for some polarization processes that would otherwise go unnoticed (chapter 8). Finally, we offer some recommendations for measuring polarization that are derived from our discussion throughout the dissertation.

# Resumen

Desde 2010, el número de recursos que aparecen en el *Google Ngram Viewer* asociados a las expresiones ‘political polarization’ y ‘political polarisation’ ha crecido enormemente. Aunque algunos autores han puesto el foco en los beneficios del aumento de la polarización política (Stavrakakis 2018), buena parte de la atención académica y mediática que ha recibido la cuestión de la polarización política ha estado centrada en sus consecuencias negativas para el funcionamiento adecuado de las instituciones democráticas (Carothers & O’Donohue 2019; McCoy & Somer 2019). La polarización puede matar la democracia, advierten Levitsky y Ziblatt (Levitsky & Ziblatt 2018), y puede ser incluso el resultado de un exceso de política y democracia (Talisse 2019), que nos vuelve más radicales (Sunstein 2017) y nos convierte en arrogantes sabelotodo (Lynch 2019). Pero independientemente del diagnóstico que se ofrezca, parece crucial esclarecer de qué hablamos exactamente cuando hablamos de una sociedad polarizada.

El término ‘polarización política’ puede utilizarse para referir a un buen número de fenómenos diferentes. Además de a los mecanismos que están relacionados con el aumento de polarización y a las consecuencias de la misma, la etiqueta a menudo se emplea para hablar de diferentes formas y tipos de polarización. Dos nociones destacadas en la literatura especializada son las de polarización ideológica y polarización afectiva. La primera es comúnmente entendida como aquella que tiene que ver con la separación en las creencias ideológicas de una población, ya sea en términos de la distancia de las opiniones en una distribución ideológica, la dispersión de las mismas, u en términos de otro parámetro utilizado para representar las creencias de una población en un espectro ideológico (Bramson et al.

2017). La polarización afectiva, por el contrario, es comúnmente concebida como aquella que tiene que ver no con las creencias ideológicas de una población, sino con sus estados afectivos hacia las personas de su propio grupo y hacia las personas del grupo ideológico opuesto (Iyengar et al. 2012).

Aunque la polarización ideológica parece tener que ver con las creencias y la polarización afectiva con los sentimientos, ambas nociones tratan de medir polarización utilizando cuestionarios de auto informe. Para saber qué piensa o siente la población, el modo habitual de medición consiste en preguntar directamente a la gente cuáles son sus estados mentales. En este sentido, dos asunciones filosóficas compartidas por ambas nociones son las siguientes. Primero, ambas nociones asumen que hay una diferencia entre los estados mentales del tipo de las creencias y los estados mentales de tipo afectivo, que explica la diferente naturaleza de ambos tipos de polarización. Segundo, ambas nociones asumen la tesis de la autoridad de la primera persona: la sinceridad de una persona garantiza la verdad de su auto atribución mental, y por eso es suficiente preguntar a la gente por sus creencias y sentimientos para saber qué es lo que creen y sienten.

Un primer problema asociado con la polarización ideológica es que su relación con las consecuencias perniciosas del aumento de polarización en una sociedad no es nada obvia. La diversidad de opiniones parece ser uno de los pilares esenciales de cualquier democracia, y sin una explicación adicional no resulta nada evidente por qué una sociedad cuyas creencias políticas estén divididas en bloques o dispersas en una distribución ideológica será necesariamente una democracia en peligro. Un segundo problema de esta noción parece ser la sublimación del punto medio del espectro ideológico: no es evidente por qué las creencias situadas en el centro de una distribución ideológica son necesariamente preferibles a aquellas ubicadas en un extremo. El aumento de la identificación con una ideología "de centro" podría ser un problema para la democracia si, por ejemplo, tales personas no estuvieran dispuestas, sistemáticamente, a coordinarse ni a convivir con personas con ideas políticas diferentes. En este sentido, la polarización ideológica no parece dar cuenta de los problemas asociados con el aumento de polarización. La polarización afectiva, por el contrario, sí consigue explicar por qué el aumento de polarización supone un problema para el buen funcionamiento de la democracia:

el aumento de los sentimientos positivos, como la simpatía, hacia las personas del propio grupo y de los sentimientos negativos, como la hostilidad, hacia las personas del grupo opuesto explican la dificultad en el consenso y la coordinación con "los otros".

A pesar de que la polarización afectiva parece estar en mejor posición que la polarización ideológica para dar cuenta de los peligros asociados con el aumento de la polarización política, esta noción también exhibe ciertas limitaciones. En primer lugar, el diagnóstico que se sigue de la polarización afectiva, tal y como se suele concebir en la literatura, es que la polarización política no tiene que ver con lo que creemos sino con lo que sentimos. Este diagnóstico no casa muy bien con la tendencia a producir discurso racional, a ofrecer argumentos que respaldan las creencias ideológicas del grupo con el que uno se identifica, que habitualmente exhibe una población polarizada. La gente no simplemente odia a sus contrincantes, sino que piensa que la evidencia les da la razón y que la otra parte está claramente equivocada. En segundo lugar, esta noción de polarización no parece poder dar cuenta de ciertos procesos de polarización en los que los sentimientos de una población permanecen estables y sin embargo aumenta el nivel de impermeabilidad hacia los argumentos esgrimidos por la parte contraria.

Además de estas dificultades, las asunciones filosóficas de ambas nociones de polarización parecen altamente cuestionables. Un buen número de experimentos (y de experiencias cotidianas) apuntan a que habitualmente fallamos cuando tratamos de identificar nuestro propio estado mental, ya sea este una creencia o un sentimiento (Schwitzgebel 2008, 2011a,b). Y tampoco está nada claro que las creencias y los sentimientos sean estados mentales radicalmente distintos con respecto a su vínculo con la acción. La primera de estas dos objeciones parece crucial para ambas nociones: nada asegura que las herramientas de medición de polarización midan con precisión lo que tratan de medir.

El objetivo de esta tesis es ofrecer una noción de polarización que no se comprometa con ambas asunciones filosóficas, evite las dificultades que presentan las dos nociones de polarización y pueda dar cuenta de algunos procesos de polarización que pasan desapercibidos para estas nociones. Para ello proponemos una serie de requisitos que, a nuestro juicio, una noción adecuada de polarización debe satisfacer. En primer lugar, esta noción debe poder explicar los efectos perniciosos para la democracia del aumento de polarización. En segundo lugar, debe ser consistente con nuestra mejor evidencia acerca de cómo nos polarizamos. En tercer lugar, no debe culpar a la gente ni explicar la cuestión en términos de irracionalidad. En cuarto lugar, debe acomodar la distinción entre decir que uno cree o siente tal cosa y que uno realmente crea o sienta tal cosa. Finalmente, debe permitir la intervención.

Este trabajo se estructura en ocho capítulos y tiene dos argumentos centrales, uno negativo y otro positivo. El argumento negativo defiende que las nociones de polarización ideológica y polarización afectiva, tal y como se entienden en la literatura, tienen una serie de limitaciones difíciles de esquivar y no miden lo que tratan de medir. El argumento positivo mantiene que las herramientas utilizadas para medir polarización afectiva en realidad miden no (solo) los sentimientos de la gente, sino las actitudes que *expresan* tener, en concreto el grado de confianza depositado en las creencias centrales del grupo ideológico con el que se identifican. Así entendida, la polarización afectiva cumple los requisitos propuestos y evita los problemas asociados con la noción.

Para cumplir nuestro objetivo, en primer lugar (capítulo 2) presentamos un estado de la cuestión de la polarización política sobre creencias trazando distinciones conceptuales en torno a tres categorías, a saber, formas generales de polarización, tipos de polarización y concepciones de polarización. Una distinción crucial aquí es la que hay entre el contenido de una creencia y el grado de creencia (Talisse 2019), es decir, la distinción entre qué cree la población y cómo lo cree. Esta distinción está en la base de dos concepciones bien distintas de la polarización sobre creencias. La polarización ideológica se centra exclusivamente en lo que la gente cree, representada por una de las acepciones recogidas en la RAE del verbo ‘polarizar’: “orientar en dos direcciones contrapuestas”. Sin embargo, hay otro modo de entender la polarización sobre creencias, basada no en lo que la gente cree, sino en cómo cree la gente lo que cree, representada por otra de las acepciones recogidas en la RAE para el mismo verbo: “concentrar la atención o el ánimo en algo”.

Posteriormente introducimos la noción de polarización afectiva tal y como habitualmente se entiende en la literatura, presentamos y discutimos tanto las limitaciones que las nociones de polarización ideológica y polarización afectiva presentan como las asunciones filosóficas que comparten, y ofrecemos el conjunto de condiciones anteriormente introducido y que una noción adecuada de polarización debería satisfacer (capítulo 3).

Para satisfacer el requisito de la evidencia sobre cómo nos polarizamos, llevamos a cabo una revisión de la literatura relevante sobre mecanismos de diversa naturaleza relacionados con el aumento de polarización y, crucialmente, discutimos la compatibilidad entre estos mecanismos y la polarización sobre creencias tanto en términos de contenido de creencia como en términos de grado de creencia (capítulo 4). La polarización ideológica, aquella que tiene que ver con los contenidos de nuestras creencias, parece no poder acomodar la evidencia, o por lo menos parece estar en peor posición que la polarización sobre creencias entendida como grado de creencia. La mejor evidencia de la que disponemos, además, parece más compatible con la idea de que la polarización es un proceso racional que con la historia irracional de la polarización (Dorst 2020).

Para satisfacer el requisito de la disanalogía entre lo que decimos que creemos y sentimos y lo que creemos y sentimos, necesitamos una teoría acerca de las adscripciones de estados mentales disposicionales que pueda acomodar la diferencia entre auto atribuirse un estado mental y expresar un estado mental, y que rechace la tesis de la autoridad de la primera persona sin violar nuestras intuiciones en un montón de casos naturales de auto atribución de creencias y otros estados mentales disposicionales. En el capítulo 5 discutimos la familia de teorías que engloba a un buen número de posiciones que de un modo u otro se comprometen con la idea de que la función de nuestras auto atribuciones mentales es la de describir un estado de cosas, y defendemos que estas teorías no pueden dar cuenta del requisito de la disanalogía.

En el capítulo 6 ofrecemos una aproximación a las atribuciones mentales basada en nuestra interpretación de un buen número de observaciones de Wittgenstein que satisface el requisito de la disanalogía. De acuerdo con esta aproximación, las atribuciones mentales no cumplen una función descriptiva, y su valor de verdad no depende de que un determinado estado de cosas, interno o externo al individuo, sea el caso, sino de la compatibilidad de su verdad con el resto de rasgos relevantes del contexto. Tener un estado mental no es una cuestión de hechos, sino una cuestión normativa. Cuando nos auto atribuimos creencias o se las atribuimos a otras personas, lo que hacemos es adquirir un conjunto de compromisos conceptuales vinculados con la acción cuya verdad dependerá de su compatibilidad con el resto de rasgos salientes del contexto. Esta aproximación abre la posibilidad de error en la atribución de estados mental.

En el capítulo 7 nos apoyamos en algunos expresivismos contemporáneos para defender que la polarización afectiva está en mejor posición que la ideológica para medir los estados mentales de la población, nuestros compromisos conceptuales especialmente vinculados con la acción. La razón principal es que las herramientas que habitualmente se emplean para medir polarización afectiva involucran lenguaje evaluativo, y a través del uso evaluativo del lenguaje expresamos nuestros compromisos, nuestras actitudes, conceptualmente ligadas con nuestra manera de ver el mundo. Es decir, a través de nuestras evaluaciones expresamos nuestros estados mentales, nuestras creencias, que a veces no coinciden con las creencias y los estados mentales que nos auto atribuimos. Crucialmente, algunas de las actitudes que expresamos, aquellas con un vínculo especial con la acción, están conceptualmente conectadas con el nivel de credibilidad que tenemos en ciertas creencias centrales del grupo ideológico con el que nos identificamos. Este movimiento nos permite reevaluar el concepto de polarización afectiva, que llamamos polarización en actitudes para distinguirlo del modo tradicional en el que se entiende el concepto, de manera que permite evitar los problemas asociados con la polarización afectiva y la tesis de la autoridad de la primera persona, y permite también satisfacer los desiderata propuestos y dar cuenta de otros procesos de polarización que de otra manera pasarían desapercibidos (capítulo 8). Finalmente ofrecemos una serie de recomendaciones para medir polarización de manera más precisa que se derivan de nuestra discusión a lo largo de todo el trabajo.

# Chapter 1

# Introduction

In the last week of February 2020, two teachers, from different schools, were arrested in Ceuta, Spain, under the suspicion of having sexually abused several of their students, 5-year-old children. Child sexual abuse is one of the most abhorrent crimes, in part because of the immense asymmetry between the victim and the perpetrator. The victims of this kind of abuse are extremely vulnerable: they are not in a position to understand what is happening to them, and the consequences are devastating. Given the particularly repugnant nature of this type of crime, it is only to be expected that a case of child sexual abuse would trigger enormous anger and aversion; even the mere suspicion that someone has perpetrated such a crime would arouse a visceral rejection. However, the very opposite happened in this case.

On February 27, 2020, immediately after the second arrest, hundreds took to the streets, including many teachers, to denounce both the supposed vulnerability of teachers when they are denounced, and the alleged existence of a 'huge mafia' dedicated to falsely accusing teachers for profit (El Pueblo 2020). During the demonstration, one of the protesters said: "If someone wants to denounce in order to make money or whatever, let them think about it. We're going to demonstrate as often as possible together with everybody else. Today for you, tomorrow for me" (our translation). "A 'Nescafé salary' –an expression referring to a prize offered by Nestlé consisting in a monthly salary for life– will not be achieved by denouncing a teacher" (our translation) said another.

At first glance, this shocking reaction is very hard to understand and explain, not only because of the nature of the crime that the teachers were accused of, but also because the day of the demonstration there was no public relevant information about the cases that could lead to suspecting about anyone in particular. The city had previously had similar cases where the accused turned out to be guilty. I was astonished when I heard about the demonstration, I didn't understand a thing –people taking to the streets to *defend* child molesters? But then I found out that there was some piece of information floating around that, unfortunately, seems to have been precisely what triggered the reaction of the demonstrators in this case: at least some, if not all, of the victims were Muslims.[1]

Despite the fact that approximately half of the population of Ceuta is Muslim, the racism and marginalization suffered by this part of the population is very high, and the situation has significantly worsened in recent years. According to the *European Islamophobia Report 2019*, the Muslim population of Ceuta "still suffers segregation, with hundreds of minors without schooling and lacking prosecutors specialized in discrimination and hate crime" (Bayrakly & Hafez 2020: 740). Anotther clue to understand the puzzle raised by this case is that the far-right Spanish political party *Vox* has received growing support over the last few years in Ceuta, and in fact won the 2019 general election in the city. The leader of this party, Santiago Abascal, has made statements such as the following one, included in the report cited above: "Islamists want to destroy Europe and Western society by celebrating the fire of Notre Dame. Take it into account before it's too late". This was not an isolated episode. For instance, *Vox* also used an image of a Hijab-wearing candidate of the left-wing party *Unidas Podemos* in Ceuta to tweet: "This twenty-year-old is a candidate for *Podemos Ceuta*. We didn't know that women's liberation consists of wearing a purple hijab". Two of the ideas that can be considered at the core of this political party's ideology, at least during its 2018-2019 campaign, were that Muslims want to invade Spain (as if being Muslim makes you

[1] In a recent work, Alba Moreno and I have discussed this case in relation with the connection between political polarization and testimonial and discursive injustices (Almagro & Moreno forthcoming)

non-Spanish) and take advantage of the country's social aids, and that numerous false accusations are only used to make a profit (see Bayrakly & Hafez 2020). It is only in this context that the demonstrators' reaction begins to make sense: given their assumptions, their behavior can be explained in terms of reasons.

It is our contention that this abhorrent and unjust case, more particularly the disproportionate reaction of demonstrators –regardless of whether the accused turn out to be guilty or innocent, can be understood as one of the pernicious consequences of the rise of political polarization. The rise of polarization, and in particular the rise of radicalism or the increase in the confidence deposited in the core beliefs of certain political identities, provides, among other pernicious outcomes, the ideal stage for this sort of case.

According to a October 2019 Pew Research study (Pew Research 2019), the level of division and animosity in the United States has deepen in relation with their already high levels of polarization previously in 2016. The trend of polarization experimented by this country in the last decades is not an isolated phenomenon. Several comparative studies have shown that the rise of polarization is taking place in a large part of contemporary democracies (Boxell et al. 2020; Carothers & O'Donohue 2019; Gidron et al. 2020; Westwood et al. 2018). In accordance with this process, there is also an increase in polarization, globally experienced, toward ethnic and racial groups (Carothers & O'Donohue 2019; Johnston et al. 2015; Mounk 2018: 166; Wojcieszak & Garrett 2018). In the case of Spain (see, for instance, Viciana et al. 2019), several national studies confirm this trend. For instance, data from the study *Opiniones y Actitudes de la Población Andaluza hacia la Inmigración* (OPIA) indicate a tendency in this line: the data from the 2019 *Encuesta OPIA VIII* shows, according to the authors of the study, that there is polarization of the attitudes of Andalusians toward immigration, and a significant growth in the group of those who strongly oppose the expansion of immigrants' social rights. In particular, this study shows that the percentage of Andalusians who perceive immigration to be one of the main problems for the population of Andalusia has doubled with respect to the previous study in 2017. Regarding the study "Attitudes toward immigration" of the *Centro de Investigaciones Sociológicas* (CIS), the data show that, in 2017, 53.3% of the Andalusian population agreed or strongly agreed with the statement that immigrants abuse free health care, compared to the 43% in 2016. And 56.3% of them see it as between quite acceptable or very acceptable to prefer to hire a Spaniard rather than an immigrant in 2017, compared to the 54.3% in 2016. These data indicate a strong rejection toward certain social groups and, although they are very high in both years, an increasing trend may be observed.

Of course, racism is not a new thing resulting from the rise of polarization. In that sense, someone might reasonably think that the demonstrators' disproportionate reaction could be simply explained by appealing to their racism, and not to the rise of polarization. This is true, but misleading: even though racism has been present in the city for a long time, the current level of polarization not only increases racism, but also makes a group of people dare to publicly display such *attitudes*, to be more confident in the core beliefs of the ideological group that they identify with. To put it in a different way, even if it is not a new thing that racists exist, perhaps it is a relatively new thing that racists feel that they have the right to blatantly express their racist attitudes in such a way, compared to what they thought they were entitled to some decades ago. Increased polarization leads to increased reliance on the core beliefs of the political group that one identifies with, and with it comes an increase in certain attitudes. It is the rise of radicalism that leads some people to publicly join such a demonstration and not only to think that denouncers are lying for profit, as a non-polarized racist would. One of the things that we will try to show in this thesis is that some conceptions of polarization obliterate this explanatorily useful distinction between polarized and non-polarized racists.

Despite the available data just mentioned, it is not exactly clear how the type of polarization behind the case introduced at the beginning –the type of polarization that has to do with *attitudes*– should be measured, nor what is it measured precisely. For instance, if you directly ask the demonstrators from our previous case whether they think that Muslim people mostly denounce to make a profit, or whether they are racists, surely their answers will be in the negative, especially if asked at a time where the level of polarization was high, but not so high as it is now (in fact, in many of the questions that appear in the CIS surveys previously mentioned, the population declares that diversity is positive, and that they are not racist at all. In the 2017 survey, nearly 74% ranked between 0 and 4, on a scale from 0 to 10, where 0 means 'not racist at all' and 10 means 'very racist'. Specifically, 40.2% placed themselves at 0). And it is not hard to think that they were answering sincerely to the questions of the study, even though their answers seem obviously false. But do they actually believe what they sincerely say that they believe? Their verbal and nonverbal behavior expresses quite the opposite. Then, on what basis do we attribute a certain state of mind to somebody? It is noteworthy that one might believe such things and, at the same time, be unwilling take to the streets, and publicly display certain attitudes. How can this type of polarization be measured then? What are we talking about when we invoke this sort of polarization? We address these and other related questions in this work.

This dissertation is an attempt to make a modest contribution to better understanding the phenomenon of polarization that endangers many contemporary democracies. In particular, we deal with the nuances of different notions of polarization as well as with the philosophical assumptions behind two prominent concepts –ideological polarization and affective polarization. Our main goal is to offer a suitable concept of polarization and some recommendations that allow us to measure polarization with greater accuracy, and in an early stage. For this purpose, we reassess the concept of affective polarization. Can philosophy offer any special theoretical tools to accomplish this task?

## 1.1. Philosophy and polarization

Polarization has been one of the phenomena that most attention has attracted, academically but also publicly, during the last decade. Scholars from different disciplines, especially from political sciences and social psychology, have been studying the causes, origins and mechanisms through which a society becomes polarized, and have analyzed the consequences of the rise of polarization for democracy. What does philosophy have to say about this issue? Is there a particular perspective that philosophy can bring to bear in addressing the polarization issues that we need to deal with?

Certainly, contributions from philosophy on this issue are increasing, and from a variety of perspectives. In an already classic paper on the epistemology of disagreement, Kelly wonders whether it is rational to respond as people usually do in situations where, after being exposed to the same mixed body of evidence, two parts that previously disagreed on certain topic, polarize (Kelly 2008). Some authors have discussed the implications of this phenomenon for different epistemological questions, such as the notion of epistemic peerhood and disagreement in general (Hallsson 2019) or the role of testimony in forming aesthetic beliefs (Robson 2014). Others have focused their philosophical analysis on another phenomenon related to the rise of polarization, which consists in deliberating with likeminded people (Arvan 2019; Olsson 2013; Talisse 2019). Some of them take not only an epistemological point of view, but also a metaphysical one (Broncano-Berrocal & Carter 2021). From virtue epistemology, and more particularly from the branch focused on epistemic vices, the issue of polarization has also been much discussed (Cassam 2019; Lynch 2019; Tanesini & Lynch 2021). For instance, Lynch has recently argued that it is the very attitude of intellectual arrogance, i.e., behaving as know-it-alls, that's behind the current political situation in many contemporary democracies (Lynch 2019, 2021).

We will occasionally deal with questions related to these ones, but they are not the center of this dissertation. Rather, we mainly approach the phenomenon of polarization from the philosophy of language and the philosophy of mind. First, in line with works aimed at distinguishing different types of polarization (Bramson et al. 2017), we carried out a conceptual analysis of different phenomena related to polarization. The term 'polarization' has been used as a catch-all word to refer to the problems that many contemporary democratic societies around the world face, as well as the mechanisms fueling polarization. Alas, the word often encompasses a wide variety of meanings, which sometimes increases the confusion about the phenomenon of political polarization itself (Bramson et al. 2017; Mason 2013, 2015, 2018; McCarty 2019). What does exactly mean to be politically polarized? Intuitively, one answer might be that our political opinions have become more extreme, and this jeopardizes democracy. This characterization looks correct. However, it is still profoundly ambiguous. This ambiguity is what partially explains the lack of agreement on some contemporary debates on polarization.[2]

In order to try to dissolve these misunderstandings, two complementary strategies might be followed. The first one revolves around the task of trying to disentangle the different phenomena usually called 'polarization'. According to this first strategy, the main task would be to *conceptually delimit* different aspects and concepts involved in the topic of polarization. That is, the task would be to disentangle the different conceptual relations established by different uses of the term 'polarization'. The second strategy consists in establishing a set of criteria that the notion of polarization should meet, and rule out those phenomena that, although related to polarization, do not meet these criteria. The main task of this second strategy, then, would be to *evaluate* different aspects and concepts of polarization in order to provide the requirements that an appropriate concept should meet.

In this dissertation we follow both strategies. Admittedly, the main objective that we follow is essentially bound to the second one: we seek to establish a set of desiderata that a concept of polarization should satisfy in order to be a suitable one, and will try to outline and vindicate a notion that meets these requirements (in line with previous work such as Bordonaba & Villanueva 2018). However, the first strategy is also followed in this dissertation insofar as we need to review the literature about polarization to put on the table first, before establishing the desiderata, the different issues related to the phenomenon that should be taken into account. In other words, the task of establishing the conditions that the notion of polarization should account for comes only after reviewing the most outstanding aspects, issues, and phenomena related to political polarization. The results of the latter task will serve as the backdrop for the former.

Conceptual clarification is not the only contribution that we try to make from philosophy here. We also hinge on issues of the philosophy of language as well as questions at the intersection of the philosophy of language and philosophy of mind. In particular, we discuss the compatibility between various *descriptivist*

[2] For the debate on whether there is general polarization in a particular country, mostly in the United States, see Abramowitz 2006, 2007, 2010; Abramowitz & Saunders 2008; Abramowitz & Stone 2006; Brewer 2005; Hetherington 2001; Hetherington & Rudolph 2015 vs. Fiorina 2017; Fiorina et al. 2008; Fiorina & Levendusky 2006; Levendusky 2009; Wolfe 1998.

positions on mental state attributions and the measurement of the states of mind that we hold, rather than those that we claim to be in, and argue that a particular nondescriptive and pragmatist account of Wittgensteinian inspiration is in a better position than the descriptivist ones to do so. Moreover, we deal with questions that are more specific to the philosophy of language, such as what sort of meaning is communicated through certain uses of language, and what theory allows us to accommodate the distinction between the commitments that we say we have and the commitments that we express we have through our use of language. This distinction is crucial both for measuring the mental states that we actually have, rather than the ones that we say we have, and our practical attitudes related to our level of confidence in certain beliefs. As racists will deny that they are so, polarized people presumably will deny that they are polarized. But polarization not only has to do with two groups of people holding different and conflicting belief contents. Some types of polarization, those that will be crucial for this dissertation, have to do with certain attitudes especially linked to action. These *affective* attitudes are related, we will argue, with the level of confidence in certain beliefs, rather than with the actual content of these beliefs. Particularly, we will explore an expressivist approach in order to accomplish this task, that is, in order to introduce some philosophical tools that enable us to discriminate between the mental states that people actually have and those that they say they have, but also between the practical attitudes that people express to have, connected with how impervious they are toward the reasons coming from the "other side". In this sense, with this dissertation we try to make a contribution to the analysis of the phenomenon of political polarization from a specific philosophical approach that has not yet been much discussed in relation with this topic.

## 1.2. *The political turn*-driven spirit

Contemporary democracy is troubled. Cassam has recently pointed out that polarization may be promoted by some political actors to advance their political agendas:

> What triggers the process of polarisation? One possibility is that it is triggered by the actions of political actors. This suggests that polarisation is a political strategy or tool that is knowingly and deliberately employed by political actors as a means of achieving their own political ends [...] Understood in this way, polarisation need not be pernicious but it often is. It can lead to authoritarianism, intolerance and disagreements over basic facts. (Cassam 2021: 213)

Cassam recognizes that polarization is not necessarily a pernicious phenomenon, but it often is. The case presented at the beginning of this introduction is just an instance of some of the pernicious consequences that it may generate. In more general terms, political polarization, when it is pernicious, "routinely weakens respect for democratic norms, corrodes basic legislative processes, undermines the nonpartisan structure of the judiciary, and fuels public disaffection with political parties. It exacerbates intolerance and discrimination, diminishes societal trust, and increases violence throughout society. Moreover, it reinforces and entrenches itself, dragging countries into a downward spiral of anger and division for which there are no easy remedies" (Carothers & O'Donohue 2019: 1-2).

So, the rise of some types of polarization might endanger democracy and create certain injustices, but it might also be promoted for political gain –by those actors for whom an electoral advantage is perceived to be linked to a polarized public opinion. Realization of this situation is the point of departure of the analysis contained in this dissertation. In other words, we start our research from the recognition that there is at least one type of polarization that poses a danger for the proper functioning of democracy, and that this type of polarization can be purposely promoted. Thus, it seems crucial to know as much as possible about this type of polarization. We need to understand how it works, how it can be detected at an early stage, and how we can intervene in it. Our theories will be better than others to the extent that they enable us to achieve these objectives.

In that sense, the spirit of this dissertation falls under what has been called *the political turn in analytic philosophy* by Pinedo and Villanueva (Pinedo & Villanueva forthcoming; see also Bordonaba et al. forthcoming). The core idea of this political turn is that our theoretical tools must be assessed paying attention to their capacity to explain and detect social injustices, as well as to intervene in order to eradicate or alleviate them. In their words, the political turn "advises us to embrace theories by their capacity to bring our focus towards hidden forms of injustice as well as for its potential to intervene on them. And [...] to reject those theories that do the opposite" (Pinedo & Villanueva forthcoming: 3). It is not enough to put our theories and theoretical tools at the service of practical concerns, it also requires to assess them in virtue of their usefulness to detect and alleviate certain injustices. This is the general framework in which this work is placed. In addition to this general framework, what other philosophical assumptions are necessary to provide the kind of concept of polarization that we offer here?

## 1.3. The pursued *river-bed*

Another starting point of this dissertation is the recognition that adopting a philosophical position might be similar to putting on some peculiar glasses that allow us to see things from a different angle, to *seeing-as*. Furthermore, we also start from the recognition that there is no such thing as the best glasses, once and for all; the preference of some glasses over others will usually depend on the goals that we pursue, along with the assumptions one is not willing to give up. In this way, in this dissertation we will try to build some specific glasses and, at the same time, to persuade you that these glasses are reasonably beneficial to look at a problem that threatens our contemporary democratic societies.

The glasses-metaphor employed, while useful, might also prompt certain erroneous associations. It is therefore worth pointing them out to avoid misunderstandings. First, it may subtly introduce the idea that there is something like *the* naked vision of the world, that is, that we can see with the naked eye how things are, and that therefore putting on some glasses is, in some sense, distorting what the world is like. Second, it may introduce the idea that changing the way you see the world is as simple as taking off one pair of glasses and putting on a different one. Both ideas are wrong. First of all, we think it is a big mistake to think that the world is in a certain way once and for all (see, for instance, Rorty 1980). Of course, this neither implies the rejection that there are things in the world, nor the rejection that through language we sometimes refer to the objects that populate our surroundings and inform about their appearance and location. The mistake is just the idea that the world is in a certain way and that our scientific endeavor just consists in discovering the way things are. Second, putting on new glasses is an extremely difficult task, as difficult as educating our dispositions and getting rid of certain habits in which we have been educated. Changing glasses, in our sense, implies reeducation, learning to see-as; it implies living in a different way.

As a result of reading this dissertation, we hope to contribute to the plausibility of some of the following aphorisms, that we take to be hard rocks, in the sense of Wittgenstein's river-bed metaphor, i.e., as unalterable parts on which our practices, the flux of the river, rest.

1. One relevant notion of polarization conceives it as an increase in the level of confidence in the core beliefs of a political identity.

   1.1 This kind of polarization has to do with the degree of belief, rather than with belief contents.

   1.2 Affective polarization can be understood in terms of credence.

   1.3 One can be polarized while being at the middle of an ideological spectrum.

   1.4 This type of polarization is highly context-sensitive.

2. Being in a dispositional mental state is having certain conceptual commitments linked to certain courses of action.

   2.1 The possibility of error exists in identifying our own mental states.

   2.2 What we sincerely say we believe or feel is not necessarily what we actually believe or feel.

   2.3 Mind is not something spooky or mysterious or describable.

   2.4 Mind and language are two faces of the same coin.

3. Through certain uses of language we express our practical commitments, especially linked to action.

    3.1 The meaning we communicate is beyond our control.

    3.2 The meaning we communicate is highly context-sensitive.

    3.3 Social identity is crucial to determinate the meaning communicated.

    3.4 Certain claims, in certain contexts, are closely related to certain ways of living.

4. Philosophy is conceptual clarification and persuasion into new ways of seeing.

    4.1 Philosophy must be resistance, a way of fighting against social injustices.

    4.2 Philosophy must be jointly developed.

    4.3 Philosophy must try to avoid postulating ontologically bizarre entities.

    4.4 Philosophy must take into account the intuitions of competent speakers.

## 1.4. The argument

The core thesis defended in this dissertation can be stated directly: the concept of affective polarization can be reassessed in order to avoid the difficulties, some of them shared by the concept of ideological polarization –which are the two most prominent concepts of polarization, that the standard understanding of it faces, and thus being able to account for processes of polarization that otherwise would pass unnoticed, such as the one mentioned above. According to the reassessment that we propose, affective polarization has to do with radicalism, i.e., with the level of confidence people have in the core beliefs of their political identity. This level of radicalism, crucially, is closely connected with people's attitudes, certain commitments especially linked to action that are expressed through the evaluative use of language.

This thesis involves two central arguments. The first one runs as follows. Ideological polarization is standardly understood as a separation in the political beliefs, in particular in the belief contents of at least two groups of people, while affective polarization is standardly understood as dealing with feelings and emotions, rather than with beliefs: the greater the negative feelings toward the out-group and the positive feelings toward the in-group, the greater affective polarization. Ideological polarization is mostly measured through direct self-report questionnaires, and affective polarization mostly through the feeling thermometer, also a direct way of self-reporting one's feelings. Thus understood, both concepts encounter several difficulties of a diverse nature that we introduce in chapter 3, but more specifically they seem to share two philosophical assumptions that are challengeable. The first philosophical assumption is that there is a sharp distinction between belief-like and desire-like mental states, which is what characterizes the main difference between them. The second one is the first-person authority thesis, the idea that the speaker's sincerity usually guarantees the truth of her mental state self-ascriptions. Since both philosophical assumptions are theoretically and empirically challengeable, both concepts of polarization encounter some problems that are hard to overcome. Moreover, these concepts, as they seem to be commonly understood, are in a bad position to account for the type of polarization behind cases such as the one introduced at the beginning of this chapter.

The second central argument supporting the main thesis of this dissertation can be summarized as follows. The tools employed to measure affective polarization, or at least most of them, include evaluative uses of language, which permit to measure not only the feelings that respondents self-report, but also their practical attitudes connected with their level of confidence in the core beliefs of the ideological identity they identify with. That's the reason why affective polarization can be found where ideological polarization is absent. To argue so, we adopt a Wittgensteinian nondescriptivist approach to mental state attributions and an expressivist approach to the evaluative use of language. The Wittgensteinian approach enables us to argue that there is a possibility of error in ascribing dispositional mental states, because they are normatively linked to certain courses of action. The expressivist approach enables us to argue that through the evaluative use of language people express their practical commitments, their attitudes. Understood in this way, affective polarization, or at least some of its types, does deal with beliefs, in particular with the degree of some of the beliefs we hold. It assumes neither first-person authority, or a sharp distinction between belief-like and desire-like mental states. Moreover, under this interpretation, affective polarization, or ”polarization in attitudes”, as we will call it to differentiate it from the standard understanding of affective polarization, can satisfy a group of conditions for a suitable concept of polarization, and can account for the process of polarization behind the case introduced at the beginning.

Besides these two central arguments, in this dissertation the reader can find other arguments supporting some theses derived from, or related to, the central thesis. For instance, along this dissertation we argue that the best available evidence on how we get to polarize is more compatible with the idea of *radicalism* than with the idea of *extremism* –radicalism has to do with the degree of belief, while extremism has to do with belief contents, and therefore with the concept of polarization in attitudes rather than with the concept of ideological polarization. Moreover, we argue in favor of the rational story of polarization: if we take into consideration all the best available evidence together, we can make good sense of the idea that becoming polarized is not the result of an irrational process, but a possible outcome of being rational in an informational environment such as ours. Also, we argue that the possibility of error in identifying our own mental states, that stem from the rejection of the first-person authority thesis, is incompatible with descriptivist approaches to the analysis of mental state attributions. Descriptivist positions cannot accommodate the distinction between the commitments one self-reports to have and the commitments one actually has. Other more promising positions that seem to be able to accommodate this distinction cannot account, however, for our different intuitions in a variety of cases of mental self-ascriptions, as we will see. We argue that a particular expressivist position, inspired by some ideas taken from Wittgenstein, can account for these things. In the next section, we offer a more detailed exposition of the general structure of this dissertation.

## 1.5. Plan of the dissertation

In chapter 2, we introduce the phenomenon of political polarization and carry out a conceptual analysis of different forms, types and understandings of polarization, focusing on the sort of polarization that has to do with political beliefs (section 2.1). First, in section 2.2, we introduce different general forms of polarization, such as belief vs. set of beliefs polarization (section 2.2.1), state vs. process polarization (section 2.2.2), elite vs. mass polarization (2.2.3), intragroup vs. intergroup polarization (section 2.2.4) and symmetric vs. asymmetric polarization (2.2.5). Then, we present some more specific types of polarization (section 2.3), in particular we introduce the concepts of platform polarization (section 2.3.1), adherence polarization (section 2.3.2) and partisan polarization (2.3.3). After that, we discuss one of the most prominent concepts of polarization, ideological polarization (section 2.4), and present nine different senses of it (section 2.4.1). In section 2.5, we differentiate two ways of approaching the study of our beliefs: in terms of content or in terms of degree of belief, and we argue that this distinction points to two understandings of polarization: extremism and radicalism. Crucially, we show that ideological polarization is conceived in terms of belief content. In section 2.6, we discuss whether political polarization is necessarily a pernicious phenomenon or, on the contrary, only some types of polarization are. Finally, we end this chapter by introducing the distinction between cognitive vs. conative polarization (section 2.7).

The aim of chapter 3 is to introduce the concept of affective polarization as it is commonly understood in the literature, to make explicit the philosophical assumptions behind the concepts of ideological polarization and affective polarization, and to propose a set of desiderata for a suitable concept of polarization. First, we present some of the problems, of a diverse nature, that the concept of ideological polarization encounters (section 3.1). Second, we introduce the concept of affective polarization as well as the tools commonly used to measure it (section 3.2), present different possible types of affective polarization (3.2.1), and discuss several problems that the concept of affective polarization, as it is commonly understood, also faces (section 3.2.2). Third, we devote section 3.3 to make explicit the philosophical assumptions behind the concepts of ideological polarization and affective polarization: the distinction between belief-like and desire-like mental states (section 3.3.1) and the first-person authority thesis (section 3.3.2). After that, we make a first attempt to challenge both philosophical assumptions (section 3.4), from a theoretical and from an empirical point of view. Finally, taking into consideration what has been discussed so far, we introduce five desiderata that a suitable concept of polarization should meet (section 3.5).

Chapter 4 is concerned with the mechanisms that foster polarization, paying special attention to whether they fit more naturally with radicalism (degree of belief) or with extremism (belief content), as well as whether they support a rational or an irrational story of the rise of polarization. We start by introducing the phenomenon according to which we get polarized after discussing or deliberating with likeminded people, which we call 'likeminded deliberation', and discuss different mechanisms involved in this phenomenon (section 4.1). Then, we introduce another phenomenon according to which two individuals who disagree on a particular issue get polarized after being exposed to a mixed body of evidence, which we call 'mixed evidence disagreement', and discuss some mechanisms and approaches to this phenomenon (section 4.2). In section 4.3, we review three psychological mechanisms related to the rise of polarization, namely: group membership, motivated reasoning and identity-protective cognition. Section 4.4 is devoted to reviewing three social mechanisms: filter bubbles, echo chambers and cybercascades. After this, we review three linguistic mechanisms also related to the rise of polarization: abstract and concrete uses of language, dogwhistles and crossed disagreements.

In chapter 5, we discuss whether descriptivist approaches, as well as Bar-On's position and a position that can reasonably be attributed to Srinivasan, can accommodate our intuitions in different cases of belief self-ascriptions. The difference between the states of mind that someone says to be in and the mental states in which she actually is, that follows from the rejection of the first-person authority thesis –one of the desiderata proposed for a suitable concept of polarization, should also be accommodated by a suitable theory of mental state ascriptions. First, we briefly introduce the difference between self-ascribing a mental state and expressing a mental state (section 5.1). Second, we introduce two varieties of descriptivist positions, and argue that they cannot accommodate the intuitions triggered by different cases of belief self-ascription (section 5.2). In section 5.3, we discuss neo-expressivism, Bar-On's approach to mental self-ascriptions. Section 5.4 is devoted to discussing a position that can reasonably be attributed to Srinivasan, based on her epistemic externalism. Finally, we also discuss certain alleged peculiar types of mental states, such as aliefs, unendorsed beliefs and in-between cases (section 5.5). We conclude that descriptivist views don't seem to be well positioned to account for our purposes in this dissertation.

The aim of chapter 6 is to introduce an approach to the mind with respect to which our concept of polarization is going to be able to meet the aforementioned desiderata. The approach holds that having a dispositional mental state is having conceptual commitments linked to certain courses of action, and that those conceptual commitments are contextually and normatively determined. We start by making a first attempt at introducing our interpretation of Wittgensteinian nondescriptivism (section 6.1). We offer an interpretation of some of Wittgenstein's insights, more specifically we introduce an sketch of the conceptual map of the psychological vocabulary and the notion of 'description' that can be traced throughout Wittgenstein's entire production (section 6.2). Then, we present a group of anti-descriptivist arguments we find in Wittgenstein's philosophy (section 6.3). After that, we devote a section to briefly discussing the relation between following a rule and our interpretation of Wittgenstein's picture of the mind (section 6.4). We also offer some indications as to the relation between our interpretation of Wittgenstein's approach to the mental and his approach to meaning (section 6.5). Finally, we present the notion of contextual authority that follows from our discussion so far in that chapter (section 6.6).

In chapter 7, we present a semantic theory, expressivism, as one that enables us to explain why some of the tools used to measure affective polarization in fact measure the state of mind that people express to be in, in particular their level of credence in certain beliefs, i.e., their level of radicalism. Expressivism can accommodate the difference between the descriptive and the evaluative, and argue that through the evaluative use of language we express our practical attitudes, those especially linked to certain courses of action. Thus, this theory is characterized by being able to accommodate the difference between the commitments we say we have and those we actually have. First, we introduce the distinction between the descriptive and the evaluative from an intuitive point of view (section 7.1) and present some tests and arguments that support this distinction (sections 7.1.1, 7.1.2 and 7.1.3). Then, we introduce expressivism (section 7.2), and we make a brief historical note about it, from classical expressivism to some more contemporary proposals (section 7.2.1). After that, we present a noninternalist expressivism, building on minimal expressivism (section 7.3) and on some Wittgensteinian insights (section 7.3.1). Finally, we devote section 7.4 to reflect on some of the contextual determinants of evaluative meaning and the difference between judgements vs. claims about the rule we think we follow.

In chapter 8, we introduce the modified concept of affective polarization, that we call polarization in attitudes, in light of the results achieved in the previous chapters (section 8.1) and discuss how this concept of polarization meets the desiderata (section 8.2). After that, we briefly discuss some of the attitudes related to the increase of polarization according to the literature on epistemic vices (section 8.3), specifically the attitudes of arrogance, dogmatism and closed-mindedness (section 8.3.1), and discuss whether they are necessarily bad attitudes and whether we are responsible for them (section 8.3.2). Moreover, we review some recent studies of affective polarization by taking into consideration the recommendations to measure polarization that follow from the discussion of this dissertation (section 8.4), and, finally, we try to sketch a design to measure polarization following these recommendations (section 8.5).

Chapter 9 is the conclusion of this dissertation. In it, we summarize what have been discussed and achieved along this work and review the conclusions reached in each chapter (section 9.1). Also, we briefly discuss a variety of mechanisms that sometimes are used as weapons to promote polarization, such as recurrent debates and crossed disagreements (section 9.2). Finally, we end this chapter by briefly reviewing and discussing some strategies to try to depolarize (section 9.3).

# Chapter 2

# Political Polarization: The Phenomenon

Consider the following fictitious but perfectly possible scenario. Imagine that during the first two months of the COVID-19 pandemic a Spanish citizen called Dereca believed that the policies adopted by the Spanish government to control the spread of the virus were insufficient. Suppose that, after discussing about it with likeminded people, maybe with her Facebook and WhatsApp friends, as well as with her Telegram groups, neighbors and closest relatives, Dereca ended up believing not just that the policies adopted by the Spanish government were insufficient, but that the development of the pandemic in Spain was in fact the government's fault. Imagine that, as a result, Dereca and many other likeminded people took to the streets to complain against the government, carrying golf clubs and Le Creuset *cocottes*, and yelling that the Spanish government deserves to be known as The Death Government. "The government must pay for its disastrous management of the situation", someone says. At the same time, imagine that also during the first two months of the COVID-19 pandemic, another Spanish citizen called Izquerri believed that the policies adopted by the Spanish government to control the spread of the virus were mostly reasonable. Suppose that, after arguing about it with likeminded people, Izquerri also became more extreme in her previous position. In this case, however, she ended up believing that the policies adopted by the government were not only mostly reasonable, but the best that could have been adopted under the circumstances. Suppose that since then Izquerri and many other likeminded people have not admitted any criticism of the Spanish government's management of the situation, and have openly complained about the behavior of Dereca and her peers. But not only that. Since then, Izquerri has treated disrespectfully all those who think differently from her. What's happened here?

It is not controversial to state that, in this fictitious scenario, Dereca and Izquerri, along with their respective groups, have become more politically polarized –both parts have become more extreme in their political beliefs. As a result, the political distance between them has increased. Thus, at first glance, what's happened in this case might count as an instance of political polarization. But what does it exactly mean to say that the political distance between Dereca's and Izquerri's groups has increased and as a result they have become more polarized?

We devote this chapter to introduce the phenomenon of political polarization, review an essential part of the literature about it, and try to separate the different concepts working under the superficial grammar of the term 'polarization', that is, the different uses of the word that can be traced in the literature about polarization. In particular, in this chapter we focus on the sort of polarization that has to do with our *beliefs*. The main aim of this chapter will be to show that the concept of ideological polarization is conceived in terms of changes in *belief contents*, and not in terms of *degree of belief*. The main thesis will be that this distinction between belief content and degree of belief points to two different understandings of polarization. To argue so, we introduce different forms, types and understandings of belief polarization, and analyze the possibilities of combination with each other. The sort of understanding of polarization that has to do with changes in the degree of belief in certain beliefs is behind the concept of polarization in attitudes that we propose along this dissertation, as a result of the reassessment of the concept of affective polarization.

In section 2.1, we briefly introduce the phenomenon of belief polarization, the type of polarization that has to do with beliefs, paying special attention to a mechanism that promotes belief polarization, a well-known phenomenon in the literature sometimes dubbed as 'group polarization', which will be further discussed in chapter 4. The aim of section 2.2 is to present various general forms of polarization, which are compatible with other more specific concepts of polarization. These general forms are particular belief polarization, set of beliefs polarization, polarization as a state, polarization as a process, elite polarization, mass polarization, intragroup polarization, intergroup polarization, symmetric polarization and asymmetric polarization. In section 2.3, we present three more specific types of polarization: platform polarization, adherence polarization and partisan polarization. In section 2.4, we discuss one of the most prominent concepts of polarization in the literature: the concept of ideological polarization. In section 2.5, we introduce two different senses of the expression 'adopting a more extreme belief', namely: extremism with respect to belief content and degree of belief, and argue that this distinction points to two different understandings of polarization, captured by the notions of extremism and radicalism. Crucially, we show that the concept of ideological polarization is essentially understood in terms of extremism, i.e., as changes in belief contents. In section 2.6, we discuss whether polarization is a benign or pernicious phenomenon. Finally, in section 2.7, we briefly introduce another type of polarization that allegedly does not have to do with beliefs, but just with feelings.

## 2.1. Divided by our beliefs: Belief polarization

Since this chapter aims to provide a first approach to the phenomenon of political polarization and to some of its general forms and concerns –at least as they are commonly understood in the literature, it might be helpful to discuss the case introduced at the beginning to try to clarify what it might mean to say that a society is polarized. Let us focus on our example.

The first thing to notice is that Dereca's and Izquerri's groups have become more extreme in their *beliefs*. That is to say, whatever it is that happened in the previous case, it seems to be somehow related to the beliefs of the people involved. But it is also important to notice that, in the example, people get polarized after *discussing with likeminded people*. It is well-documented that deliberation with likeminded people seems to promote polarization. The study of this phenomenon, however, does not concern what polarization is, but how we get polarized. For this reason, we will consider this second issue further in chapter 4, where we review and discuss different mechanisms that seem to fuel polarization. Nevertheless, given the relevance of this phenomenon in the literature on polarization, it seems necessary to start by offering a few brief insights about it.

The study of what happens within a group of likeminded people when they deliberate on a particular issue –a phenomenon sometimes called 'group polarization' to emphasize the group character of the phenomenon (see Broncano-Berrocal & Carter 2021) and other times called 'belief polarization' to emphasize its link with the beliefs of the group (see Talisse 2019)– is one of the most prominent mechanisms discussed in relation to political polarization (Aikin & Talisse 2020; Breton & Dalmazzone 2002; Talisse 2019; Sunstein 2002, 2009, 2017). This phenomenon consists in the tendency of members of a group of likeminded people to hold more extreme beliefs and positions than the ones they started with after discussing with each other (see Brown 1985: 203-226).

The term 'group polarization' was coined by Moscovici and Zavalloni in the late 1960s (Moscovici & Zavalloni 1969; see also Myers & Lamm 1976 : 603), once it was acknowledged that the tendency to adopt more extreme beliefs as a result of discussing in group is a fundamental group decision-making process. Thus, the so-called *group polarization hypothesis*: "The average postgroup response will tend to be more extreme in the same direction as the average of the pregroup responses" (Myers & Lamm 1976: 603).[1] This mechanism seemed at odds with one of the main findings established in the literature until then, supported by the work of Gordon Allport and other scholars in the 1920s and 1930s: we tend to avoid expressing extreme opinions in social situations,[2] and group consensus

[1] This is a general phenomenon that has its origins in what is known as "risky shift". In 1961, James Stoner observed that group decisions were riskier than the average of initial individual decisions, and this tendency was called risky shift. Group polarization is an extension of these studies where, however, group extremism is not reduced to riskier decision making.

[2] This finding can actually be compatible with group polarization if the groups of people in which each pattern is present are significantly different (for example likeminded vs. non likeminded groups).

represents the average opinion of individuals. How can individuals acquire more extreme beliefs after group discussion if they tend to moderate their views while discussing, and consensus seemingly crystalizes around average opinions?

The tendency of the members of a group to adopt more extreme beliefs along the lines of the initial ones occurs in homogeneous contexts of deliberation. One of the earliest empirical studies regarding this mechanism was conducted on a group of 140 male students from a Parisian lycée, ages 18-19. The authors conducted three experiments in which the subjects were separated into groups of 4 participants. They were given an attitude scale form with some items to be filled out individually, then asked to discuss and reach agreement on each item, and finally they were given again the same scale form to rate each item individually once more. In the first experiment, the participants' beliefs toward General de Gaulle (e.g., "De Gaulle is too old to carry out such a difficult political task") were measured. The second one measured the participants' beliefs toward the Americans (e.g., "American economic aid is always used for political pressure"). The third experiment measured the same items that were tested in the first one but this time asking participants to evaluate how favorable each item toward de Gaulle was regardless of whether they agreed with the items or not, with the goal of measuring the results "dealing this time not with opinions but with 'objective' judgments" (Moscovici & Zavalloni 1969: 129). Each item was evaluated on a Likert scale from -3 (strongly disagree) to +3 (strongly agree). The results showed that, in all three experiments, the group consensus and post-consensus rates were significantly higher than the initial individual ones. They called this the 'polarization effect'.

This outcome has been succesfully replicated through a large number of studies (see Brown 1985; Sunstein 2002, 2009). Another well-known study tested 256 students' beliefs, at three Michigan high schools, regarding eight racial items (e.g., "Some people recently have saying that 'white racism' is basically responsible for conditions in which African Americans live in American cities. Others disagree. How do you feel?").[3] Participants' prejudices toward African Americans were

[3] Note that this item asks about feelings, which are the main feature of some tools used to measure affective polarization. This will be relevant in the following chapter.

tested first during a psychology class. In a subsequent session, their individual responses to the eight items were collected, and then participants were assembled into likeminded groups based on the results of the previous prejudices test. Next, they were asked to discuss the eight items regarding race in the United States. The 'polarization effect' replicated. The prejudiced participants became more convinced that 'white racism' was not responsible for the disadvantages suffered by African Americans after deliberating with the in-group people. The same, in the opposite direction, occurred with the non-prejudiced participants (Myers & Bishop 1970).

This effect is not limited to particular periods, cultures, nations or issues (Myers & Lamm 1976; Sunstein 2009: 3, 18-19; Talisse 2019: 102; Broncano-Berrocal & Carter 2021). Sunstein calls 'enclave deliberation' to the process of deliberating among likeminded people who mostly talk and live in isolated enclaves (Sunstein 2002: 177, Sunstein 2017).

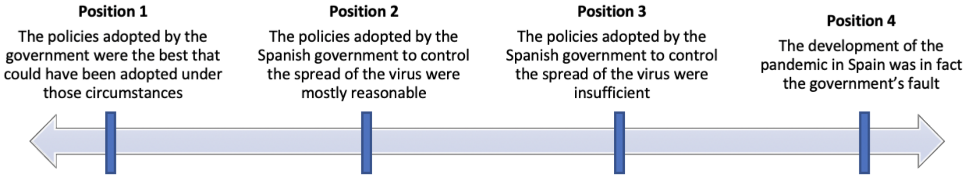
This mechanism seems crucial to understand how polarization arises, although it is not the only one that we must focus on. But, in any case, whatever happened to Dereca and Izquerri in the case presented at the beginning, it seemingly had to do with their beliefs and with deliberating with likeminded people.

## 2.2. Polarization is said of many things: General forms of polarization

In this section, we review and discuss different general forms of political polarization that have been pointed out in the literature. More specifically, we review the dichotomies of particular belief vs. set of beliefs, state vs. process, elite vs. mass polarization, intragroup vs. intergroup polarization, and symmetric vs. asymmetric polarization. All these dichotomies, as we will see, are compatible with different types of belief polarization. The phenomenon of political polarization is commonly linked to public opinion, that is, to people's political beliefs. That's the reason why we start by offering the map of concepts related to belief polarization. One of them, a particular understanding of polarization –radicalism, will be crucial for our notion of polarization in attitudes.

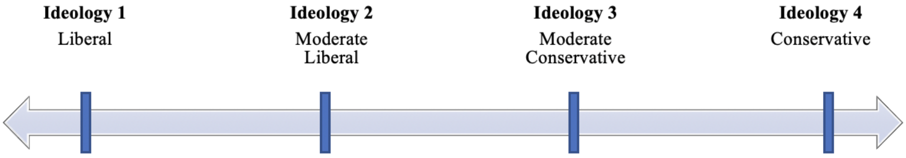
### 2.2.1. Particular beliefs and set of beliefs

The first thing to notice is that two groups of people can be polarized over a particular ideological belief or a set of beliefs. So, there can be polarization at least about two different items, namely: *particular ideological beliefs*, and *broad ideological differences*. For an initial glimpse of the matter, let us point out that people's beliefs can become more extreme just about one particular political issue. As an example, consider the four following options on the Spanish government's management of the pandemic. Position 1: The policies adopted by the government were the best that could have been adopted under those circumstances. Position 2: The policies adopted by the Spanish government to control the spread of the virus were mostly reasonable. Position 3: The policies adopted by the Spanish government to control the spread of the virus were insufficient. Position 4: The development of the pandemic in Spain was in fact the government's fault. The following image offers a visual representation of these possibilities.


Position 1
The policies adopted by the government were the best that could have been adopted under those circumstances
Position 2
The policies adopted by the Spanish government to control the spread of the virus were mostly reasonable
Position 3
The policies adopted by the Spanish government to control the spread of the virus were insufficient
Position 4
The development of the pandemic in Spain was in fact the government's fault

According to a particular understanding of what polarization means, polarization has to do with belief contents, and that we will call 'extremism' (see section 2.4) –people's opinions about the Spanish government's management of the coronavirus pandemic are more polarized to the extent that they are grouped closer to positions 1 and 4 rather than positions 2 and 3, respectively. Thus conceived, polarization has to do with the distance between the four positions. No matter what 'extremism' and 'distance' mean for now, the point here is that since this process of division could happen in a way that people's views on other political issues remain stable, then there could be polarization just on *a particular issue*. In other words, a society can be polarized over its beliefs about the Spanish government's management of the pandemic and not over other issues.

In addition, two contending groups can be polarized over their general orientation to politics, and not only over a particular matter. That is, instead of being polarized just about the Spanish government's management of the pandemic, people can be polarized about their broad ideological orientation. To see that, suppose we have the following four types of ideology. Ideology 1: Liberal. Ideology 2: Moderate liberal. Ideology 3: Moderate conservative. Ideology 4: Conservative.



Again, and according to what we called above 'extremism', people's ideologies are more polarized to the extent that they are grouped closer to ideologies 1 and 4 rather than positions 2 and 3. The distance between the groups of beliefs marks the level of polarization. Thus, a society can be polarized to the extent that about half of the population is liberal and the other half is conservative. Insofar as the terms ‘liberal’ and ‘conservative’ are two umbrella labels for a set of particular belief contents, it is only to be expected that such polarization will manifest itself in the division on different particular political issues.

In the current context, it is hard to find out a Western democracy polarized over just a single issue. However, this is not only a conceptual possibility, but an historical one (see Fiorina 2017). But nowadays, there is a high probability that if somebody has a certain political belief, then she will endorse a particular set of beliefs. As we will see, this is an essential feature of radicalism: people's high level of credence in the core beliefs of the political group that they identify with makes them to endorse an additional set of beliefs.

### 2.2.2. State polarization and process polarization

A second point that it is important to make in relation to our clarification task is that, when we said that Dereca's and Izquerri's groups had become more polarized in our example, the word 'become' can mean at least two different things. On the one hand, it can mean that people's political beliefs are *already* extreme. That is, people's political beliefs are extreme right now. The idea would be, for example, that after measuring the opinions of a population on some given topics at a particular time, the results show that their opinions are extreme with respect to these topics at *that particular time.* DiMaggio and other scholars call this approach 'polarization as a *state*' (DiMaggio et al. 1996). On the other hand, the word 'become' can also mean that people's political beliefs are *undergoing* a process of polarization. That is, people's opinions are gradually becoming more extreme, and the process itself *has not yet come to an end.* This second sense is called 'polarization as a *process*'. becoming polarized can be understood both as a *state* and as a *process* (DiMaggio et al. 1996). The essential difference between them, we think, just lies in the point of view from which the phenomenon is approached: synchronic approaches to polarization understand it as a state, while diachronic approaches conceive it as a process.

The sort of concept of polarization that we pursue in this dissertation must be an operational one. In particular, it must enable us to intervene as soon as possible, by detecting the rise of polarization in an early stage. Thus, our approach to polarization here will be one more diachronic and dynamic than synchronic and static. In this sense, this distinction will be relevant.

### 2.2.3. Elite polarization and mass polarization

Another important distinction has to do with the extension of the expressions 'Dereca's group' and 'Izquerri's group', in particular with the fact that their extension is not fully specified. With it, we can refer to different sets of people. Polarization is sometimes discussed and measured among political elites, which can include just party officials, or policy intellectuals and activists as well. When the study of polarization is focused on political elites, it is often called *elite polar-*

*ization* (McCarty 2019: 13). However, since political elites could not be entirely representative of a population, it could be that despite there being polarization among the elite, citizens' public opinion remains distributed along a normal line (Maravall 1981; González & Bouza 2009: ch. 5). It might also be that, at least conceptually, the general public is polarized and the political elite is not. Therefore, in order to know whether a country is polarized or is undergoing a process of polarization, it is necessary to measure polarization in the whole population. When the study of polarization is focused on the masses, it is often called *mass polarization* (McCarty 2019: 13). This distinction is crucial. Among the most intense debates about polarization that have taken place in recent decades is the question of whether a country –in particular the United States– is polarized on both levels (Hetherington et al. 2016; Sides et al. 2018 vs. Abramowitz & Webster 2017; Iyengar et al. 2012).[4] While there is a strong consensus on the existence of elite polarization, scholars and political pundits disagree on whether there is also mass polarization (see, for instance, Hetherington 2009; McCarty 2019: 50-54).

The type of polarization that seems to put democracy in danger is the one that mobilizes citizens, the sort of polarization that makes people become impervious to the reasons coming from the "other side". It is only the unwillingness of citizens to listen to the other side and to coordinate with them that makes certain political parties obtain political gain. The type of polarization we pursue in this work must be conceived in terms of mass polarization.

### 2.2.4. Intragroup polarization and intergroup polarization

It is normally assumed that polarization implies that something happens regarding at least two opposing groups. The rise of political polarization has to do with the increase of the political distance between two –or more– political groups. As we will see in the next section, this distance can increase symmetrically or asymmetrically. Polarization can increase because the division between at least two contending groups arises from the extremization of the beliefs of both

[4] Despite the study of polarization have been extensively tied to the analyses of the United States situation, there is growing literature on polarization in relation to other countries (see, for instance, Gidron et al. 2020 and Carothers & O'Donohue 2019)

groups or simply because one of them experiment a change. However, the point that we want to stress here is another slightly different. As we have seen at the beginning of the chapter, polarization might increase as a result of a process of deliberation within a group of likeminded people. In this sense, polarization can be analyzed simply by focusing on what happens inside one group. When likeminded people discuss about certain issues and as a result they all get polarized, we can talk about *intragroup polarization*. However, when polarization is analyzed by focusing on what occurs regarding more than one group, we talk about *intergroup polarization*.[5] The distinction between intragroup polarization and intergroup polarization, as we will see, can be conceived not only in terms of belief contents, but also in terms of degree of belief.

### 2.2.5. Symmetric polarization and asymmetric polarization

As we have seen, when we say that people's beliefs become polarized, we may refer to the political elite or to the general population. In this section, we will make another point concerning the people who can be polarized. In addition to the fact that polarization can occur among political elites and the whole population, polarization can occur *symmetrically* or *asymmetrically* (Grossmann & Hopkings 2016; McCarty 2019: 42). Since political polarization is a measure of the political distance between political opponents, no matter how this distance is exactly conceived, the distance can increase due to a shift experimented only by one group or by both.

For instance, there is certain consensus among scholars that the increasing distance between Democrats and Republicans in the United States during last decades is primarily an outcome of the Republican Party's shift to the right (see, for instance, Hacker & Pierson 2005; Mann & Ornstein 2012; Theriault 2013; Hare

[5] Even though it is conceptually possible to talk about polarization just within one group of likeminded people, it is important to note that the concept of polarization is essentially relational. If a society homogeneously believe that p, and after a while the population homogeneously increases their confidence in such a belief, then we won't talk about polarization, although technically it would be a case of intragroup polarization. Polarization requires the existence of another group of people with opposing beliefs.

et al. 2012). In this sense, it can be stated that political polarization in the United States is asymmetrical.[6]

To develop this point a little bit further, let us consider, for instance, the political context of Poland as it is described in (Fomina 2019: 126). In Poland, the right-wing party PiS (*Prawo i Sprawiedliwość*) and its adepts are ideologically very cohesive and politically mobilized. At the same time, the opposition is very fragmented and just reacts to the government's policies and rhetoric. According to this rough characterization, two different senses of asymmetrical polarization can be distinguished. On the one hand, it can be said that, in terms of dispersion of beliefs (see section 2.4), in Poland the opposition is more polarized than PiS supporters to the extent that their positions are more fragmented than those held by PiS.

On the other hand, it can be said that, in terms of commitments to their own perspective (see section 2.5), i.e., their degree of belief, PiS adherents are more polarized than the opposition to the extent that they are much more mobilized and cohesive, which means that they are more ardent supporters of their perspective than those belonging to the opposition. As we will argue through this dissertation (especially in chapters 3 and 8), one of the most pernicious types of political polarization has more to do with becoming impervious to the reasons of the opposing groups –which is linked to political identity and the degree of belief in the core ideas of the group one identifies with– than with extremism. For the time being, however, it is sufficient to point out that polarization can be symmetrical or asymmetrical to the extent that the increase in the division between at least two groups arises simply from the extremization of the beliefs of one group or the extremization of the beliefs of both groups.

Finally, we want to make another remark. According to some, allegedly surprising findings, lack of support to democracy, and even hostility to it, is strongest in the center of the ideological spectrum (Adler 2018). This outcome replicates

[6] However, the evidence supporting this claim is very mixed, and it is not fully clear what this kind of shift to the right means. But there are many studies supporting the idea that conservatives tend to polarize mor than liberals (see, for instance, (Bail et al. 2018; Heltzel & Laurin 2020; Westfall et al. 2015).

across several Western democracies. As the author of the study notes, “Respondents at the center of the political spectrum are the least supportive of democracy, least committed to its institutions, and most supportive of authoritarianism” (Adler 2018: 2). We can talk of the change in the beliefs of a group of people located at the middle of an ideological distribution. This situation counts also as an instance of asymmetric polarization to the extent that polarization does not involve a change experimented by two groups, but only by one of them.

## 2.3. Polarization is said of many things: Types of belief polarization

Besides the general forms of polarization presented above, we can distinguish other more specific types of polarization, which are compatible with those previously introduced. Let us unpack three types of polarization: platform, adherence, and partisan polarization.

### 2.3.1. Platform polarization

The first type of polarization we want to introduce here has to do with the way in which a set of issues is treated by two contending groups. Specifically, this type of polarization can be understood as the ideological distance between the platforms of competing political parties, which Talisse calls *platform polarization* (Talisse 2019: 98). The idea is that when platform polarization is on the rise, the proposals of two political parties on a set of issues strongly diverge, and then the political middle ground between people from both groups disappears. For instance, if we return to our example, we can see that Dereca’s and Izquerri’s groups might have opposing views not only on how the government of Spain has handled the pandemic, but also on their views on abortion, freedom of speech, inclusive language, gender equality, financial policies, the monarchy, etc. As the number of issues for which the platforms of each group have opposing positions increases, platform polarization increases.

Note that this type of polarization is compatible with the general forms introduced in the previous section. Platform polarization can increase because two political parties offer different proposal for one topic or for a set of issues. It can be analyzed at a particular time or during the time. Platform polarization can increase simply because the political elite of different political parties offer different proposals on certain topics, or because the supporters of those parties endorse the proposals. This type of polarization can be analyzed just by focusing on the proposals offered by one group, or by taking into account the proposals offered by both groups. Finally, it might also be the case that only one group changes their proposals on certain topics.

### 2.3.2. Adherence polarization

A second type of polarization, different from the previous one, is what we are going to call *adherent polarization*. The general idea is that there is adherent polarization when the vast majority of the population is divided at least in two political parties, either in terms of votes or in terms of party identification. For instance, in the United States, there would be adherent polarization if there were a transfer from political independents to the Democrats and Republicans. Similarly in Spain, if there were a transfer from nonpartisan people –or from people belonging to more moderate parties, whatever that means– to *Unidas Podemos* and *Partido Popular*, there would be an increase in adherent polarization. Conceptually, this type of polarization is also compatible with the general forms of polarization. Fiorina calls this type of polarization 'partisan polarization' (Fiorina 2017).

### 2.3.3. Partisan polarization

The third type of polarization that we want to distinguish is *partisan polarization*, understood as partisan ideological uniformity (Talisse 2019: 98-99), i.e., a scenario in which the policy preferences and ideology self-identification of the members of at least two political parties have become more sharply *aligned* (see, for instance, Bishop 2008; Fiorina 2017: 44-49). This type of polarization is sometimes called as *sorting* and *party sorting*, and some authors distinguish between two different forms of them: issue-based sorting and social-based sorting (see, for instance, Mason 2018). The former has to do with an alignment regarding the issue positions held within a political identity, while the latter has to do with an homogeneity of social identities, such as geographical, religious, racial, etc. This type of polarization can also be understood as a *mechanism* that facilitates the rise of polarization, at least to the extent that it promotes situations where each person is mostly exposed to discuss only with likeminded people.

Partisan polarization is a radically different type of polarization from adherent polarization. To see why, it might be useful to think a bit about the variety of members that might compose two political parties. Let us assume, for the sake of explanation, that, at a particular time, among the electorate of *Unidas Podemos*, 50% of its members are liberals, 20% are conservatives, and 30% are in the middle. Furthermore, suppose that 80% are for abortion, and 20% are against it. Imagine now that, after a few years, all members of *Unidas Podemos* become liberal and pro-abortion. Then, a process of (issue-based) partisan polarization has taken place: this party is now more internally homogeneous. Note, however, that partisan polarization and adherence polarization are logically independent processes. A political party's electorate can become more ideologically homogeneous –i.e., partisan polarized– without increasing the number of voters or people self-identified with this party (see Fiorina 2017: 46-47); it may even decrease the number of voters, or people who identify with the party, and still partisan polarization can take place. Again, this type of polarization is compatible with the general forms of polarization pointed out above.

## 2.4. A more traditional type of polarization: Ideological polarization

Beyond the previous three types of polarization, there is a more traditional, notion of polarization in the literature: the concept of *ideological polarization*. According to this concept, roughly put, polarization has to do with the representation of the beliefs of a population in an ideological distribution, either on the basis of their proximity to the poles or on other parameters.[7]

Two intuitive approaches to this concept of polarization are the following. First, polarization becomes greater as the division of opinions that differentiate at least two well differentiated positions in an ideological space augments. That is, the greater the cluster of people in two blocks or positions well differentiated from each other, the greater the level of polarization. Second, polarization rises as the dispersion between different opinions in a given distribution augments. That is, the more diverse and separate the particular opinions of a population along an ideological distribution, the greater the level of polarization. The former is known as *bimodality*, while the latter is known as *dispersion*.

> **Dispersion**: Public opinion on an issue can be characterized as polarized to the extent that opinions are diverse, "far apart" in content, and relatively balanced between ends of the opinion spectrum. (DiMaggio et al. 1996: 694)

> **Bimodality**: Public opinion is also polarized insofar as people with different positions on an issue cluster into separated camps, with locations between the two modal positions sparsely occupied. (DiMaggio et al. 1996: 694)

Thus, ideological polarization can be conceived in two senses. The first one is what these authors called 'dispersion': people's beliefs on an issue can be considered polarized to the extent that those beliefs are diverse, come apart, and are relatively balanced between the extremes of a given ideological spectrum. In this sense, the greater the likelihood that two randomly selected respondents differ

[7] This concept of polarization has received many different labels, such as 'political polarization' (Sartori 1976), 'issue position polarization' (Mason 2013: 141, 2015), 'policy-based division' (Iyengar et al. 2012), 'political preferences' (Fiorina & Abrams 2008), 'opinion polarization' (DiMaggio et al. 1996) and 'self-reported or self-described ideology' (Gentzkow 2016).

in their beliefs, the greater the polarization. The second dimension is what they called 'bimodality': beliefs on an issue can be considered as polarized to the extent that people cluster into separate positions, with few people occupying the positions between them. Note that one of the main differences between dispersion and bimodality is that dispersion is just about individual's beliefs, while bimodality is about groups of people with a similar position. Bimodality is the common conception of polarization (see, for instance, Fiorina 2017; Hetherington 2009; Sartori 1976), i.e., as two groups of people that move away from each other. When bimodality is also linked to the tendency toward the extremes of an ideological distribution, then the resultant type of polarization is equivalent to what has been called 'extremism' (see, for instance, Sunstein 2017), and that can be defined as follows:

> **Extremism**: If, at t1, agents X1...Xn and Y1...Yn respectively hold conflicting attitudes A1 and A2, then their attitudes polarize if, at t2, X1...Xn and Y1...Yn respectively hold attitudes A3 and A4, where A3 and A4 are attitudes situated more near of each pole in a given ideological spectrum.

Despite the fact that the idea of bimodality mainly comes from DiMaggio and his colleagues, they indeed offered a more complex definition of polarization, a multidimensional one. In addition to dispersion and bimodality, they distinguish two other dimensions of political polarization: *constraint* and *consolidation.*

> **Constraint**: The extent to which opinions on any one item in an opinion domain (a set of thematically related issues) are associated with opinion on any other.[8] (DiMaggio et al. 1996: 696)

[8] Note that constraint is extensionally equivalent to partisan polarization, at least to issue-based alignment.

> **Consolidation**: The greater the differences across multiple social indicators (e.g. gender, race, occupation, age, income, etc.), the greater the degree of opinion polarization between two groups. (DiMaggio et al. 1996: 698)

The idea is that polarization is not just the result of how apart and dispersed two groups of people are, or how diverse are the opinions of a population, but also how they are internally shaped. If two groups of people have substantial differences between them, but are also internally heterogeneous, then they are less inter-group polarized than if they are more internally homogeneous, both in ideological and social terms. Although the authors acknowledge that increases on different polarization dimensions indicate polarization of different kinds, they argue that political polarization entails joint increase of dispersion, bimodality, constraint, and consolidation (DiMaggio et al. 1996: 699).[9] In that sense, they offer a more complex, multidimensional notion of polarization, which distinguishes between four dimensions.

### 2.4.1. Labyrinth of paths: Senses of ideological polarization

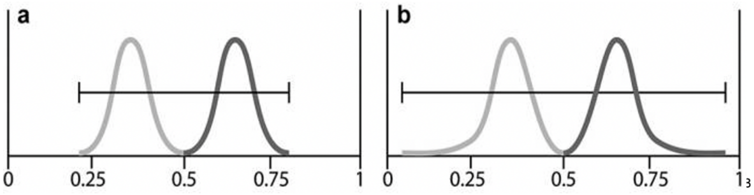
Recently, some authors have delved into the task of clarifying different senses in which we can say that a society is polarized, understanding polarization as a distribution of beliefs in an ideological spectrum. In this sense, they have developed the task initiated by DiMaggio and his colleagues of differentiating several relevant dimensions in order to determine whether a society is polarized. In particular, Bramson and other scholars (2017) have distinguished nine different senses in which ideological polarization can be understood and measured.[10] In what follows, we briefly summarize them. The first four senses are focus on observable features from the whole population. The rest are senses of polarization

[9] For a debate about this multidimensional approach to polarization see Mouw & Sobel 2001 and DiMaggio et al. 1996. For a trimodality approach to polarization see Downey & Huffman 2001.

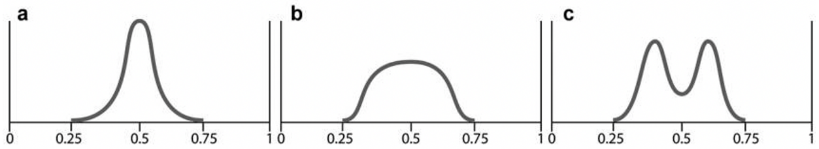
[10] Actually, they explicitly say that from their conceptual work follows that there are nine senses of polarization in general (Bramson et al. 2017: 117), but inasmuch as these senses have to do with the distribution of belief contents, these nine senses can be seen as different senses of ideological polarization.

concerning one or several groups.

- **Spread**: This first sense of polarization that they distinguish simply focuses on the breadth of opinions on a spectrum: the more the extent or length of the opinions in a distribution, the greater the polarization. In this sense, the key to measuring polarization is just attending to how apart are the most distant opinions. The notion of spread captures one of the aspects of what DiMaggio and his colleagues call 'dispersion'. In the figure below,[11] diagram b represents greater polarization in the sense of spread.

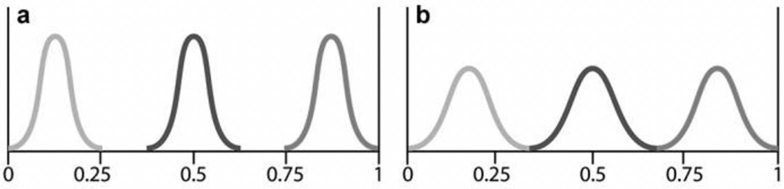


- **Dispersion**: This second sense measures polarization in terms of statistical dispersion, that is, in terms of the overall shape of the distribution of beliefs and opinions, and not only in terms of how apart the most distant opinions are. In this sense, two societies can have the same spread of opinions and different dispersion inasmuch as, although the opinions are equally extreme in both, they differ in the number of people who support one or another opinion, and therefore they have different overall shapes of the distribution. This concept also seems to fit with what DiMaggio et al. 1996 call 'dispersion', and with adherent polarization in belief content. In the figure below, diagram c represents greater polarization in the sense of dispersion.

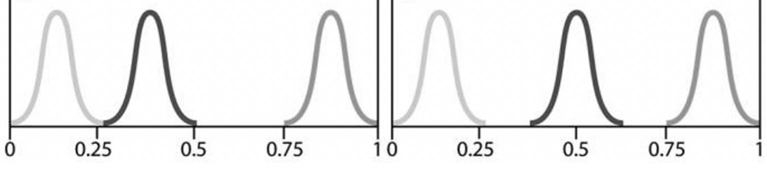


[11] All diagrams come from Bramson et al. 2017.

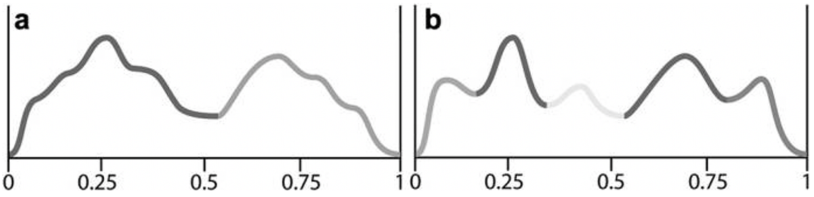
- **Coverage**: The focus on the empty space of an ideological spectrum can also indicate polarization in a different sense. That's what Bramson and his colleagues call coverage. The idea is that the narrower the opinion bands, the more empty space there will be, and therefore less diversity of opinions available. In the figure below, the diagram a represents greater polarization in the sense of coverage.



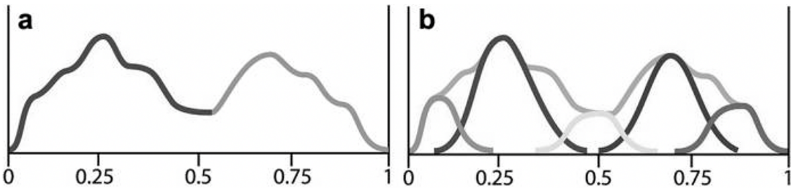
- **Regionalization**: Similar to coverage, regionalization measures polarization in terms of the empty space in a distribution. However, instead of focusing on the total amount of empty space, it measures polarization by focusing on how many regions are empty. In other words, two distributions can have the same total amount of empty space and, nevertheless, one of them can have two empty regions and the other only one region. In this case, the distribution with two empty regions indicates more polarization than the other distribution in terms of regionalization. See the figure below: diagram b represents greater polarization in the sense of regionalization, but a and b have the same coverage.



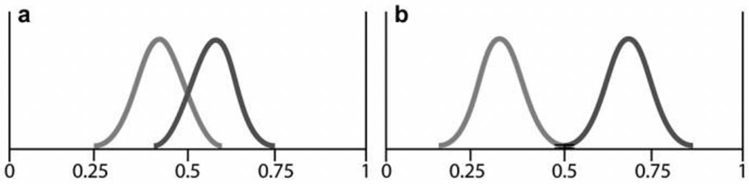
- **Community Fracturing**: This sense of polarization focuses on groups rather than concrete opinions. In particular, it measures polarization in terms of the degree to which the population can be broken into subpopulations or groups. The idea is that a society is more polarized the more fragmented the population's beliefs in different groups. In the figure below, the diagram b shows greater polarization in the sense of community fracturing. Note that in the representation b there are five groups, while in the distribution a there are just two.



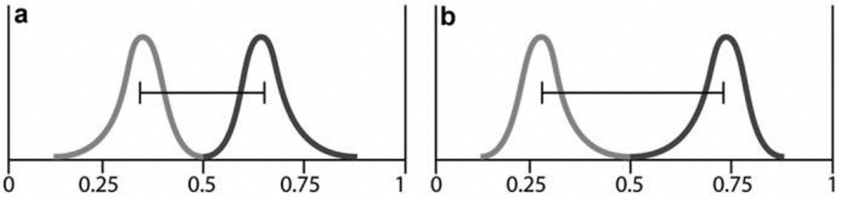
But a given subpopulation or group can also be fragmented if it is categorized in a particular way. That is, we can observe more community fracturing if we conceive the groups endogenously instead of exogenously. See the figure below: diagram b shows greater polarization in this sense of community fracturing because it distinguishes more subgroups into the two general groups differentiated in the distribution a.



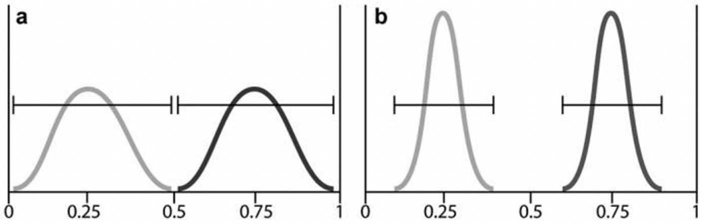
- **Distinctness**: This concept of polarization measures polarization in terms of the degree to which two belief groups are separated from each other. The idea is that the more distinct two groups are, the greater the polarization, regardless of whether they are more or less apart in the spectrum; for this conception it only matters the separation between the shape of both groups. This sense fits with what DiMaggio et al. 1996 call 'bimodality'. In the figure below, the diagram b represents greater polarization in the sense of distinctness.



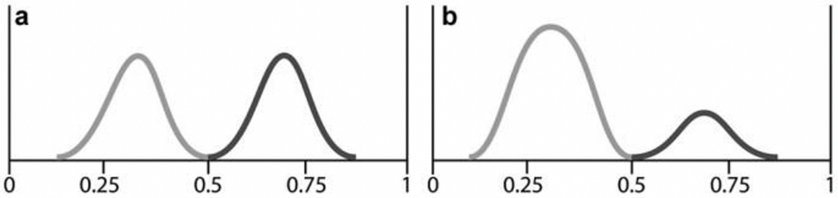
- **Group Divergence**: In contrast to distinctness, group divergence does not measure polarization by focusing on how far apart the groups are from each other, but how far apart the characteristic ideas of each group are. According to Bramson and others (2017), this sense of polarization also fits the definition of 'dispersion' in DiMaggio et al. 1996, especially when combined with an assumption of bimodality (Bramson et al. 2017: 125). In the figure below, the diagram b shows greater polarization in the sense of group divergence.



- **Group Consensus**: Group consensus polarization measures polarization in terms of how concentrated a group's beliefs are on the group's central ideology. The idea is that the more homogeneous are the views within the groups, the greater the polarization between them. In the figure below, the diagram b represents greater polarization in the sense of group consensus.



- **Size Party**: Finally, the last sense of polarization distinguished by the authors, i.e., size party polarization, measures polarization focusing on the number of people holding opposing sets of beliefs. The idea is that two societies in which there are two groups of opposing opinions, the polarization will be greater the more equal the number of people holding both. In the figure below, diagram a shows more polarization in the sense of size party.



As we have seen, ideological polarization can be measured in many different ways depending on whether we focus on one parameter or another of a given distribution of opinions or beliefs. Bramson and his colleagues have done an excellent conceptual work by distinguishing all these different senses. However, what we are interested in stressing here is precisely the features all these different notions share. All of these senses of polarization count as ideological polarization insofar as they measure polarization by attending to how people's beliefs are *distributed in an ideological spectrum*, commonly measure through *self-report questionnaires.* Hence, regardless of the particular sense adopted, ideological polarization can be defined as the concept that measures polarization through self-report questionnaires and by attending to the distribution of the beliefs represented in a given ideological spectrum. As DiMaggio and others says, "Polarization is not noisy incivility in political exchange; although the two things may be associated empirically, polarization refers to the extent of disagreement, not to the ways in which disagreement is expressed" (DiMaggio et al. 1996: 692).

## 2.5. Tools in a tool-box: Content, degree and commitment

Let's take a brief look back. So far, we have introduced general forms and types of polarization, and we have further discussed the traditional concept of ideological polarization. All these concepts of polarization have to do, or are compatible, with polarization of beliefs, specifically with the trend of adopting more extreme beliefs. But, what does the expression 'adopting more extreme beliefs' mean? We devote this section to showing that it might mean at least two different things that should not be conflated. Let's try to introduce them by discussing our main example.

Recall that, in our example, Dereca initially believed that *the policies adopted by the Spanish government to control the spread of the virus were insufficient* and, after a while, she ended up believing that *the development of the pandemic in Spain was in fact the government's fault.* Izquerri, for her part, initially believed that *the policies adopted by the Spanish government to control the spread of the virus were mostly reasonable* and, after a while, she ended up believing that *the policies adopted by the government were the best that could have been adopted under those circumstances.* According to this change, both Dereca and Izquerri have adopted more extreme beliefs, but in a particular sense: they now have beliefs with more extreme *contents.*

The traditional understanding of polarization, as we have seen in section 2.4, essentially emerges from the notion provided by DiMaggio and his colleagues (DiMaggio et al. 1996), and assumes this sense of adopting an extreme belief: it is the shifting of contents of beliefs within an ideological distribution that determines the level of polarization, under their different senses. Ideological polarization, under its different senses, is conceived in terms of belief contents.

However, there is a different way of construing the distance that is the mark of political polarization. In particular, the extremity of a belief can refer not to the content of a belief and its location in an ideological spectrum, but also to the *degree of belief* (Talisse 2019; see also Aikin & Talisse 2020: 34). To see this, suppose that Dereca and Izquerri didn't change their initial beliefs in terms of contents after discussing with likeminded people. Instead, imagine that they started to believe the same contents but with much more confidence. As a result, suppose that Dereca, although she still believes that the policies adopted by the Spanish government to control the spread of the virus were insufficient, experiences a change in her attitudes. For instance, she grabs her newly purchased Le Creuset *cocotte*, which she was not going to use because the help takes care of that, and together other people takes to the streets to complain against the government. Izquerri, for her part, still believes that the policies adopted by the Spanish government to control the spread of the virus were mostly reasonable, but now she is willing to treat disrespectfully anyone that thinks otherwise, and is no longer open to consider any opinion that is not similar to her own. In this second case, Dereca and Izquerri have adopted a more extreme belief not in terms of its content, but in terms of its *degree of belief*.

In his book *Overdoing Democracy: Why We Must Put Politics in Its Place*, Robert Talisse distinguishes a third and more subtle category, introduced to emphasize one of the possible outcomes of a situation of belief polarization that is different from the mere shift in the content of a belief and the mere shift in the degree of belief (Talisse 2019: 106-110). According to Talisse, deliberation with likeminded people can lead to the "adoption of a successor belief that is more extreme in content than its antecedent, and it also involves an intensification of the believer's *commitment* to his or her perspective" (Talisse 2019: 109, our emphasis). Deliberation with people that think alike, says Talisse, can lead both to a shift in the content of a belief and in the degree of certain beliefs, in a way that makes us more ardent supporters of a position: it increases our commitment to our own point of view. Talisse's motivation for introducing this third element is to distinguish the simple degree of belief from becoming more ardent devotees of our point of view, which also involves a shift in the belief contents. To show this, Talisse discusses a well-known experiment in the literature of deliberation with likeminded people (Myers 1975) –in which pro-feminist and chauvinist participants end up with more extreme beliefs after discussing with likeminded people –to point out the following difference. It is not the same to believe with high confidence that a woman should be as free as a man than to become a more ardent feminist (Talisse

2019: 109). The second involves more than simply increasing confidence in a particular belief. Hence, although he links it to the shift in the content of a belief, we can define this third category simply as the increased commitment to our own perspective. In our example, the idea would be that Dereca might change her degree of belief without becoming an ardent opponent to the government, and Izquerri might change her degree of belief without treating those who do not think like her with disdain.

Before leaving this discussion, we want to introduce an idea that will be of some importance in this dissertation. It is noteworthy that the issues shaping our political identity shift over time. For instance, some particular positions on the territorial organization in Spain are more politically and ideologically salient today than they were during the 1990s –e.g., believing that Spain should be a single territory controlled by the central state or that Spain should be divided into independent autonomous states are characteristic of two salient ideological identities nowadays. Therefore, a shift in the degree of belief that Spain should be a single territory controlled by the central state is closely related to a particular ideology, and to certain political parties. To the extent that a particular belief is more salient for an ideological identity, the variation in the degree of belief will be associated with an increase or decrease in the commitment to a particular ideological perspective.

In short, there are three ways in which people's political beliefs can become more extreme. Beliefs can become more extreme in their content, in the degree of belief, and in the level of commitment to the perspective for which this belief is a core one. But since the difference between the last two might be simply that the belief in which the level of confidence has increased is or not a core one to a political identity, we can reduce them to belief content and degree of belief.

### 2.5.1. Two understandings of belief polarization: Extremism and radicalism

The issues canvassed above regarding the general forms of polarization and types of polarization, together with the distinction between belief content and degree of belief, lead to some distinctions that are important to keep in mind when thinking about the meaning of 'polarization'. Recall that six dichotomies have been pointed out regarding polarization. We have distinguished between particular belief vs. a set of beliefs, state vs. process, elite polarization vs. mass polarization, intragroup vs. intergroup, symmetric vs. asymmetric, and, finally, content vs. degree. Moreover, we have introduced four types of polarization, three of which were: platform, adherent and partisan polarization. At first glance, it seems that these six dualities together with these three types of polarization simply point to conceptually combinable features that a group of polarized people might exhibit. For instance, two or more particular groups holding opposing beliefs may experience polarization among the political elite (elite) or among the population (mass), can be analyzed at a given time (state) or in terms of the trend of the belief change (process), can be analyzed in terms of what happens within a group (intragroup) or between groups (intergroups), can become extreme in a group (asymmetrical) or in more than one (symmetrical), can be polarized in terms of the content (belief content) or in terms of the level of confidence in them (degree of belief), can be polarized in terms of the amount of proposals in which both parties diverge (platform), can be polarized regarding the number of people identified with each group (adherence), and the groups can become more internally aligned (partisan).

However, note that the belief content vs. degree of belief distinction is not just a particular feature that a type of polarization can exhibit. Rather, this distinction entails two specific and different *understandings* of polarization. Regardless of whether polarization is symmetrical or asymmetrical, measured synchronously or diachronically, present among the elite or among the masses, etc., it remains to be determined what the essence of polarization, so to speak, is. It is our contention that the distinction between belief content and degree of belief aims to two different conceptions of the ideological distance that characterizes political polarization. This ideological distance, which can be characterized by all the above-mentioned forms, either consists in distance in the contents of beliefs or in the degree of beliefs.[12] To be clear, we distinguish between *general forms of polariza-*

[12] As we will see, these two types of ideological distance correspond to two different types of

*tion*, *types of polarization* and *conceptions* or *understandings of polarization*. The distinction between belief content and degree of belief corresponds to the latter, and can be grasped by the distinction between extremism and radicalism:

> **Radicalism**: At t1, agents X1...Xn and Y1...Yn respectively hold conflicting attitudes A1 and A2. Their attitudes polarize if and only if, at t2, X1...Xn and Y1...Yn give more credibility to their respective attitudes.

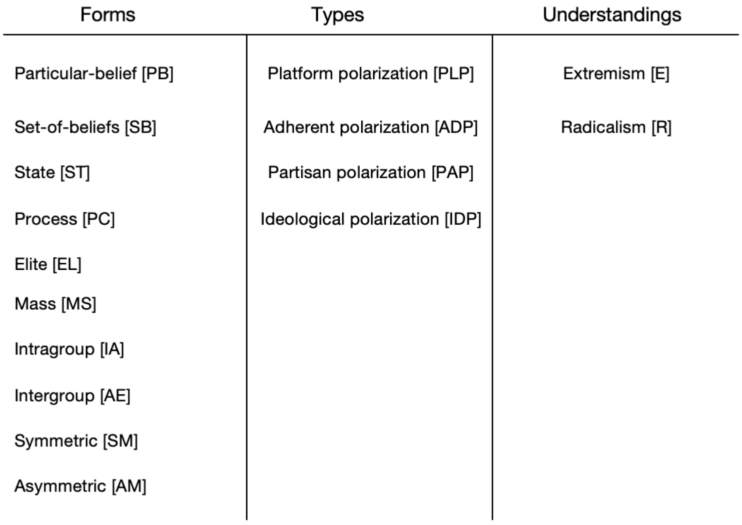
> **Extremism**: At t1, agents X1...Xn and Y1...Yn respectively hold conflicting attitudes A1 and A2. Their attitudes polarize if and only if, at t2, X1...Xn and Y1...Yn respectively hold attitudes A3 and A4, where A3 and A4 are attitudes which are situated closer to opposing poles of a given ideological spectrum.

Thus, the only difference between radicalism and extremism is that radicalism consists in a shift in the degree of belief –credence– or level of confidence, while extremism consists in a change of belief content. It is important to highlight again that the concept of ideological polarization, unlike those of platform, adherence and partisan, can only be understood as different versions of extremism, insofar as they all concern belief contents, rather than credence.

In sum, all types of polarization are compatible with all general forms of polarization. All general forms of polarization are also compatible with all polarization understandings. But not all types of polarization are compatible with both ways of understanding the notion. In particular, the concept of ideological polarization is not compatible with polarization understood as a change in the degree of belief, i.e., with radicalism. The following table offers a visual view of it.[13]

polarization: ideological polarization and polarization in attitudes.

[13] Radicalism is incompatible with ideological polarization. Except for this, both understandings of polarization are compatible with each type and form. Then, the possibilities of combination are



| Forms | Types | Understandings |
|---|---|---|
| Particular-belief [PB] | Platform polarization [PLP] | Extremism [E] |
| Set-of-beliefs [SB] | Adherent polarization [ADP] | Radicalism [R] |
| State [ST] | Partisan polarization [PAP] | |
| Process [PC] | Ideological polarization [IDP] | |
| Elite [EL] | | |
| Mass [MS] | | |
| Intragroup [IA] | | |
| Intergroup [AE] | | |
| Symmetric [SM] | | |
| Asymmetric [AM] | | |

The reader has probably noticed that an important notion of polarization is missing here: *affective polarization*. The reason is that, allegedly, this type of polarization does not has to do with beliefs, but just with feelings. We introduce it in section 2.7 and develop it in the next chapter. As we have advanced in the introduction, we will argue that affective polarization, reassessed as we do in this dissertation, indeed has to do with beliefs. In particular, with radicalism.

## 2.6. Benign and pernicious polarization

The word 'political polarization' seems to have a negative taste: when we describe a population as politically polarized, we typically thereby express our disapproval and worry about it. So, the following question arises: is it tautological

the following:

E + PLP + PB or SB + ST or PC + EL o MS + IA or AE + SM or AM
E + ADP + PB or SB + ST or PC + EL o MS + IA or AE + SM or AM
E + PAP + PB or SB + ST or PC + EL o MS + IA or AE + SM or AM
E + IDP + PB or SB + ST or PC + EL o MS + IA or AE + SM or AM
R + PLP + PB or SB + ST or PC + EL or MS + IA or AE + SM or AM
R + ADP + PB or SB + ST or PC + EL or MS + IA or AE + SM or AM
R + PAP + PB or SB + ST or PC + EL or MS + IA or AE + SM or AM

to say that polarization undermines democracy? Can the extremization of our political beliefs be a good thing for democracy? Providing the right answer, we think, requires proceeding carefully.

Some scholars have pointed out that political polarization can be beneficial to democracy in different ways. Democracy inherently values the existence of real alternatives, and even the ideological distance promoted by polarization (Stavrakakis 2018). That is, diversity of opinions, and respect for all of them, is a fundamental value of any democracy under certain comprehension of what 'democracy' means, and political polarization, understood in terms of the dispersion of belief contents in a distribution, contributes to widening the range of available options. In doing so, the argument proceeds, polarization can help to disrupt the status quo by offering new options that otherwise would not be available, which helps to overcome conformism. And since conformity can be an obstacle to moral and social progress, polarization can be positive to democracy (see McCarty 2019: 19). In fact, the American Political Science Association (APSA) issued a report in the 1950s advocating for a more polarized two-party political system that would make it easier for people to choose and to identify the reasons why they should vote for one political group over another (APSA 1950).

Furthermore, it can be argued that polarization can improve political participation and facilitate party self-identification. For instance, Abramowitz (Abramowitz 2010: 33) argues that polarization, understood as partisan polarization, produces a more engaged public and heightened political participation.[14] In this respect, Levendusky argues that sorting also simplifies the available alternatives to the voters by making them pretty clear (Levendusky 2009: 138-141), and enhance electoral accountability as well, since parties are less able to hide their positions in a fog of ambiguity (see also Blankenhorn 2015; Fiorina 2017: 79-85; McCarty 2019: 19-20).

Note that the benefits of polarization outlined above crucially depend on the concept of polarization adopted. If polarization is understood as an increase in the dispersion of belief contents in a distribution, we can derive certain benefits from it. On the other hand, if it is understood as an increase in certain aspects

[14] Although Levendusky and Fiorina have contested this approach by claiming that the data do not support their view (see Fiorina 2017: 80).

related to the homogeneity of certain groups, it can produce other benefits as a consequence. Therefore, one important conclusion we want to underline here is that whether polarization can be beneficial or detrimental depends on the type of polarization that we are talking about, and the specific context that we are discussing.

As we will see, in order to explain some of the dangerous outcomes usually associated with a high level of polarization, we have to focus on a particular type of polarization. More specifically, we have to focus on the sort of polarization that has to do with how impervious to the reasons of our political adversaries we become when we have a certain level of confidence in the beliefs that make up our political identity at a particular time (Sunstein 2017; Lynch 2019; Talisse 2019; Bordonaba & Villanueva 2018), i.e., a particular version of radicalism. Again, it is not our contention that the negative or positive outcomes of polarization are exclusively linked to the type of polarization in question. Even when it is understood as radicalism, polarization can sometimes be benign and serve as a strategy of resistance against certain harmful situations (see Pinedo & Villanueva forthcoming). We would sometimes hope, for example, that the belief of some progressives on the benefits of public education was held with a higher credence, so that clear policies about it could be expected of them when they hold power. In the following chapter we will say something more in this line.

Moreover, it is important to note that assessing polarization as benign or pernicious involves an evaluation, while the rest of the general forms considered above, e.g., symmetric or asymmetric, do not with them such a valence because they mostly involve descriptive information (see chapter 7 for the distinction between descriptive and evaluative information). But, even though every type of polarization can be deemed as benign or pernicious depending on the specific context and the specific goal, the notion of radicalism seems to be more relevant in order to explain the sort of situations that certain contemporary democracies face and that put them in risk.

In the next paragraphs, we summarize some pernicious consequences of polarization, the ones that the notion we will favor must explain.[15] One of the nega-

[15] Some authors call ‘severe polarization’ the kind of polarization that has pernicious effects (see

tive consequences of polarization is that it increases distrust in public institutions and in government, which has consequences at different levels. On the governance level, polarization can lead to policy gridlock. According to Hetherington and Rudolph, polarization diminishes *political trust*, which they define as "a feeling that people have about government based on their perceptions of its performance relative to their expectations of how it ought to perform" (Hetherington & Rudolph 2015: 35). That is, political trust is an index of people's feelings toward the government. When political trust falls, it is more difficult to build consensus and aversive reactions to the opposite side increase. The resulting scenario makes citizens less likely to seek different perspectives on controversial topics (Valentino et al. 2008). Therefore, polarization leads to an increased difficulty in coordination and cooperation, and corrodes the proper functioning of democratic institutions (Levitsky & Ziblatt 2018; McCoy & Somer 2019).

Moreover, political polarization also diminishes people's tolerance for certain issues and certain groups of people and exacerbates discrimination and violence. For instance, polarized people tend to discriminate against partisan opponents in economic transactions (Carlin & Love 2018; McConnell et al. 2018).

Consequences for ethnic minorities and for populations forced to leave their home countries are one of the greatest challenges and one of the tensest issues that contemporary Western societies face (Mastro 2015; Johnston et al. 2015; Mounk 2018). It has been shown that polarization increases national sentiment and identity, and this leads to an increase in the problems and difficulties faced by certain disenfranchised groups (see Wojcieszak & Garrett 2018).

The following quote serves to summarize some pernicious consequences of political polarization: "it routinely weakens respect for democratic norms, corrodes basic legislative processes, undermines the nonpartisan structure of the judiciary, and fuels public disaffection with political parties. It exacerbates intolerance and discrimination, diminishes societal trust, and increases violence throughout society. Moreover, it reinforces and entrenches itself, dragging countries into a downward spiral of anger and division for which there are no easy remedies" (Carothers & O'Donohue 2019: 1-2). The concept of polarization that

McCoy & Somer 2019; Carothers & O'Donohue 2019).

we pursue along this dissertation must be able to explain these negative consequences.

## 2.7. Polarization is said with respect to different mental states: Cognitive vs. conative polarization

To finish this chapter, we want to introduce a different type of polarization, which will be further discussed in the next chapter. Let's return to our initial example. In section 2.1, we pointed out that Dereca's and Izquerri's groups have become more extreme in their *beliefs*, and then we scrutinized what might mean ending up with more extreme beliefs. However, it is important to note that, according to the recent literature, certain types of polarization may not be about beliefs, but about other −non-cognitive− attitudes. In the last few decades, it has been emphasized that people can be polarized in terms of their feelings and attitudes toward adversaries. In this regard, Dereca and Izquerri might end up polarized not just in what they believe, but also in how they feel.

Despite the numerous labels used to refer to one type of polarization or another, there is some consensus that the kind of polarization that has to do with feelings and prejudices is a phenomenon of a different nature, called *affective polarization*. Affective polarization is usually defined as the tendency to dislike the other side over and above their policy preferences (Talisse 2019). On the other hand, the most prominent form of polarization that has to do with political beliefs is, as we have seen above, *ideological polarization*. The distinction is built, it is assumed, from the allegedly different nature of the two kinds of mental states that each concept deals with. Ideological polarization has to do with belief-like states, which are deemed cognitive ones, while affective polarization has to do with feelings and other action-oriented states, which are deemed conative mental states (see section 3.3). So it is assumed that ideological polarization tools are used to measure people's beliefs, while affective polarization tools are used to measure people's feelings. Thus, it could theoretically happen that people were not polarized in their beliefs but in their feelings.

Throughout this dissertation we will argue that conceiving ideological polarization and affective polarization as dealing with states of mind of a different nature is a mistake: certainly, each of these concepts point to a range of different situations and use different tools to measure polarization, but we will argue that they do not deal with mental states which are different in their nature. According to our diagnosis, the difference between them is that ideological polarization is conceived as dealing with belief-content, while affective polarization deals with credence. But that is a result that lies far ahead of us at this point, and before we get out of the rabbit-hole, we need to fall into it. Since it is assumed that both concepts point to phenomena of a different nature, we will call them here *cognitive polarization* and *conative polarization* to emphasize the nature of the phenomenon they allegedly measure and keep the labels 'ideological polarization' and 'affective polarization' for two different ways of conceiving and measuring political polarization. For now, and keeping the spirit of presenting a general picture of polarization to use it as the starting point of our discussion, it is sufficient to stress that there may be polarization not just about what people believe, but also about what people feel. Then, if Dereca's group and Izquerri's group dislike each other more than they did some years ago, then conative polarization has increased.

## 2.8. Conclusion

In this chapter, we have presented the state of the art of the phenomenon of political polarization, paying special attention to polarization of beliefs. In particular, we have presented some of the main aspects, forms, concepts and understandings of polarization that have been discussed in the literature. We have tried to separate the notions that point to general forms related to political polarization, from the notions that point to different concepts and different understandings of polarization.

What kind of polarization have Izquerri's and Dereca's groups experienced? It seems to be neither platform polarization, nor adherence polarization, nor partisan polarization. There could be some of this, but as we have presented the case there are not enough elements for it to be explained in terms of these types of polarization. On the other hand, it could be seen as a case of ideological polarization insofar as it involves changes in belief contents. But this explanation would be incomplete. A crucial point of our example had to do with the *practical attitudes* of Izquerri's and Dereca's groups. Part of what is implied by the increase in polarization in that case was a shift in people's attitudes, a change in what they are willing to do.

One of the main conclusions of this chapter is that the distinction between belief content and degree of belief points to two different understandings of polarization. In particular, belief content is essentially related to the framework of extremism, unfolded in the different senses of polarization that we have grouped under the label 'ideological polarization'. Degree of belief, on the other hand, is essentially tied to the framework of radicalism. Different types of polarization can be understood under both frameworks, with the obvious exception of two types: ideological polarization, and affective polarization. Are there also different sorts of radicalism? It will be our contention that certain forms of affective polarization involve radicalism, and therefore degree of belief, but not all forms of affective polarization entail radicalism.

How to measure political polarization? What problems do the concepts of ideological polarization and affective polarization pose? Do different approaches try to measure the same thing? Our answer throughout this dissertation will be that attempts to measure ideological and affective polarization both point to the same phenomenon, namely: a particularly dangerous process that jeopardizes democracy. But attempts to measure affective polarization do it with greater precision, because they usually take into account different parameters. To reach this conclusion we first need to make explicit, in the next chapter, the theoretical and philosophical assumptions behind both concepts of polarization, and establish the desiderata that an adequate notion of political polarization must meet.

# Chapter 3

# Affective Polarization, Philosophical Assumptions and Desiderata

In 2016, the Madrid city government, led by the then Mayor Manuela Carmena from *Unidas Podemos*, made a few changes in the annual parade celebrated on January the 5th. In particular, they decided to eliminate the VIP seats –in order to have more space for people with mobility limitations and thus enabling them to enjoy the parade, to prohibit the participation of animals –a demand repeated by groups fighting for animal rights, and there was no longer a white person with his face paint black to simulate the third wise man, Balthazar. Moreover, for the first time, more than one woman embodied a page and a magician. Even one of the "Three wise men" was a woman. Finally, they also changed the characters' clothes for others less orthodox, designed by the stylist Jorge Dutor.

These changes triggered a really huge storm of comments loaded with hostility and anger, mainly through social media, toward Manuela Carmena and her team. For instance, a headline in the newspaper *El Mundo* read "Manuela Carmena imposes an ethnic and un-Christmas-like parade", and some people called the parade "the three wise hipster men". But one of the things that generated a huge reaction was one tweet from Cayetana Álvarez de Toledo, now an exdeputy of *Partido Popular*, who wrote: "My 6-year-old daughter: Mom, Gaspar's suit is not real. I will never forgive you, Manuela Carmena. Never". This tweet unleashed the popular Twitter hashtag #NoTeLoPerdonareJamasCarmena, where people from Cayetana's side dropped all kinds of hostile messages toward Carmena and her government.

Arguably, this case counts as an instance of affective polarization, to the extent that there is a group of people displaying increasingly hostile attitudes and negative feelings toward the opposite group, while no particular change of beliefs seems to be at play. It is reasonable to think that in a less affectively polarized context, the same facts would not trigger such negative feelings and attitudes.

It is commonly assumed that affective polarization, contrary to ideological polarization, does not have to do with beliefs, but with feelings. In order for affective polarization to be said to increase between two groups from t1 to t2, no belief change is needed, the change mainly concerns the feelings of both groups. That's the reason why the difference between ideological and affective polarization is commonly taken to be one grounded on the kind of mental states that each one deals with: ideological polarization has to do with belief-like mental states, which are not linked with action –in a sense to be later specified, while affective polarization has to do with desire-like mental states, more motivational, practically-oriented, mental states.

This is usually the story, but it is not the only possible one. We agree that ideological polarization and affective polarization are two very different types of polarization. However, we don't think that the crucial difference between them has to do with the nature of the kind of mental states that each one actually deals with. In particular, in later chapters we will argue that affective polarization, or at least some types of affective polarization, mostly deals with the degree of confidence that people *express* to have in certain beliefs, more specifically their confidence in some core beliefs of their political identity –i.e., radicalism, while ideological polarization, as we have seen in the previous chapter, conceives polarization in terms of changes in the contents of beliefs. Sometimes, the tools used to measure ideological polarization actually measure the attitudes that respondents *express*, in particular those regarding their level of confidence in certain beliefs.

But, sometimes also, some tools employed to measure affective polarization do not actually measure what people express, but just the mental states that people *self-ascribe*. That's one of the reasons why we need to clarify what is exactly at issue with each type of polarization, and how it can be measured.

But before arguing about that, we first need to discuss a little bit more some difficulties, of a diverse nature, faced by the concept of ideological polarization (section 3.1), and to introduce the concept of affective polarization as it is commonly understood, as well as to discuss some difficulties that it also faces (section 3.2). Next, we will make explicit two of the philosophical assumptions behind the concepts of ideological and affective polarization such as they seem to be commonly understood in the literature (section 3.3). These philosophical assumptions are 1) that there is a sharp distinction between belief-like, and pro and con feeling-like mental states, which are those that each one deals with, and that 2) there is a default first-person authority concerning one's mental states, given the way in which both types of polarization are often measured. After that, we make a first attempt to challenge these assumptions (section 3.4), and finally we offer a set of conditions that a concept of polarization should meet in order to be a suitable one (section 3.5).

As in the previous chapter, we will use the case introduced at the beginning to discuss part of the contents developed here. At the end of this chapter we also summarize the lines we have to follow to reassess the concept of affective polarization in a way that can satisfy the desiderata, and we'll call it *polarization in attitudes* to distinguish it from the traditional way of understanding affective polarization.

## 3.1. Too low, terrain! Ideological polarization and its limits

As we have seen in the previous chapter, the concept of ideological polarization involves different senses of polarization, all of them dealing with belief contents. In this sense, they all are in a sense versions of extremism. Does ideological polarization, in its different versions, offer an accurate way to measure the pernicious level of polarization in a population? In order to address this question, in this section we will review some problems related to the concept of ideological polarization. More specifically, we unfold the array of limitations concerning ideological polarization by distinguishing between those that stand against the tools used to measure –in particular against direct self-reporting questionnaires– and those that stand against conceiving polarization in terms of shifts in belief content. Taken together, these limitations weaken the capacity of ideological polarization, under all its forms, to adequately reflect the level of polarization that seems to endanger the well-functioning of democracy.

A first concern is related to how successful direct self-report questionnaires are to measure what people actually believe about some ideological issues. This worry implies realizing that there is a variety of reasons why people might mostly rate in the middle of the offered options when asked about political and ideological questions. One of the reasons for that, usually pointed out in the literature (Converse 1964; Bullock et al. 2015), hinges on the fact that part of the population pays little attention to politics, which can be seen as a form of "rational ignorance" (Downs 1957) since they obtain little benefits from caring about it (see Hetherington & Rudolph 2015: 17). With this in mind, it is to be expected that most of the survey responses, at least those of the less politically informed, will cluster in the middle of the scale used, or will be about the most "moderate" options (Palfrey & Poole 1987; McCarty 2019: 189).

Of course, strategies to overcome this difficulty can be found in the literature. Some authors have proposed to filter the responses of those less politically informed in order to achieve a more accurate representation of what a population actually thinks about a political issue (see, for instance, Hare et al. 2015; other ways to try to avoid this limitation can be seen in Adler 2018: 3). However, one could be a supporter of a certain political option without knowing much about the issue, so this first solution does not seem to be a very effective one.

At any rate, there are more strong reasons why people may mostly respond by choosing the middle option when asked about political issues. For instance, it has been proven that, even among the well informed, most people usually avoid choosing an "extreme" option in their answers. The reason is that extreme choices are deemed as radical, which has a negative connotation. It appears that there are not many people that think of themselves as holding ”extreme” or ”radical” positions, even when they do (see Hetherington & Rudolph 2015: 19). Of course, sometimes people explicitly choose an extreme option when answering to a self-report questionnaire, but this is more likely to happen when the level of polarization on a particular issue is actually significantly high and not in many other circumstances where there is also polarization. When this occurs, such a response often indicates a high degree of attachment to a certain group, an expression of political bigotry, and not so much a well-informed preference. As Lauka, McCoy and Firat argue, ideological positions are sometimes just markers or “empty signifiers” to citizens, and they do not really believe what they say they do (Lauka et al. 2018).

To complicate things further, it is important to realize that political and ideological questions are often complex and abstract ones, and sometimes are misunderstood or not fully understood even by those who are politically well-informed. Hence, in order to appear as a nonradical person, or to avoid taking risks and sounding ignorant, it makes sense to think that most people tend to choose the most “moderate” options even when that option does not actually represent what they think about the issue. As Hetherington and Weiler put it, “Surveys tend to depress dispersion because respondents, especially the ill-informed, tend to choose the midpoint of survey items regardless of their true preferences (if such preferences can be gleaned at all)” (Hetherington & Weiler 2009: 20). Thus, people’s responses may be largely in the middle for many different reasons, preventing us from accurately grasping what they actually believe, and how polarized they actually are in terms of the content of their beliefs. Besides, some complex issues can be liked in the abstract, but such responses may not reflect respondents’ actual opinions on the matter. As McCarty puts it: “Overly simple questions such as “do you want a tax cut?” or “do you want to reduce inequality?” are also not very informative. Many people like these things in the abstract, but in real life tax cuts and inequality reduction come with trade-offs” (McCarty 2019: 188).

In addition, the order of the questions in a survey can also prompt different answers, and thus respondents might provide the answer that they think fits well with interviewer's opinion, or with their preferred ideological side (see McCarty 2019: 187-190). Another concern regarding the survey design to measure ideological polarization has to do with the problem of deciding which answers will be deemed as the most and the least extreme. Of course, we have an intuition about which options are more or less extreme, normally based on how restrictive they are, in a direction or another. For example, the opinion that *abortion should be always illegal* seems more extreme than thinking that *abortion should be illegal but with some exceptions*, because the first option is more restrictive than the second. In this sense, the poles of a distribution can be construed in relation to how restrictive some ideas are. However, sometimes it is difficult to decide which options will be situated in an end of the ideological spectrum and which ones closer to the middle. To see an example, consider the following question from the *Centro de Investigaciones Sociológicas* of Spain (CIS):

> Who do you believe is contributing most to the current political tension?
> 1) The media and journalists.
> 2) Politicians and political parties.
> 3) Entrepreneurs and economic powers.
> 4) All equally.
> 5) Other.
> 6) Don't know.
> 7) No answer.

Which of these options is more extreme than the rest? In what sense? Why could a shift from one of them to another be seen as a symptom of polarization in terms of belief contents? Certainly, there are some more or less convincing intuitive answers. But it seems that none of them can be based on some options being more or less restrictive than others.

A third more serious concern, related to the very idea of polarization behind the concept of ideological polarization, is that there seems to be a sublimation of the middle point. The options located in the middle of a distribution are commonly seen as preferable to the options located near the poles. But according to some recent findings, centrists are sometimes the most critical and openly hostile to democracy, and most likely to support authoritarianism (Adler 2018). Moreover, the intuition that the middle point is preferable to the extremes tends to be blurred when some particular cases are taken into consideration. For example, is it less preferable to think that abortion should be legal without any restriction than that abortion should be legal but with some restrictions? Why should the opinion 'abortion should be legal but with some restrictions' be a preferable –more moderate– one?

Although we don't share it, an intuitive response in favor of it might be that inasmuch as two groups of opinion are situated in the extremes of a distribution, they have less common ground, and therefore it will be more difficult for them to reach agreement and coordination. This intuitive answer draws on the idea that disagreement poses a difficulty to reach coordination. Some authors support this idea by arguing that agreement on the standards from which we make judgments is essential to coordination (Egan 2010: 260; Marques & García-Carpintero 2014; Plunkett 2015; Sundell 2016). However, the matter is more complicated than it might appear. To see this, we only need to consider that most of us know people that, despite holding beliefs located far from our own about a large array of topics, we love discussing with. But not only that. In fact, most of us learned many things from people with different views, we educate our standards by discussing with them, and frequently we reach surprisingly stable agreements. Exposing oneself to the reasons of those who do not think like us –with the exception of certain particular situations where in fact exposing to the others' reasons may polarize us (see sections 4.2 and 4.5) or damage our knowledge (Almagro et al. forthcoming; Fricker 2007)– seems important for our training in knowledge-acquiring practices. That's exactly what is advocated by those who maintain that disagreement is a source of learning and coordination (see citealtBordonaba2017), or by those who defend that acknowledging that the truth or falsity of our evaluations is relative to our standards is what enable us to explain moral progress (Pérez-Navarro 2019, 2021).

Avoiding to engage, as a norm, with the arguments of those who do not think like us, far from being beneficial, is one of the most powerful polarizing mechanisms (see section 4.2), and can also lead us to develop certain epistemic vices (Cassam 2019: ch. 5; Medina 2013). Hence, it's not really intuitive to think that occupying positions located far apart in a distribution is enough to promoting conflict, as it is not that it is preferable to hold similar opinions. In fact, the very opposite can be argued: the farther away the extreme views in a society in terms of belief contents are, the more diversity of opinions will be available in that society. As Aikin and Talisse put it, "political disagreement among political equals is central to democracy" (Aikin & Talisse 2020: 2). Thinking that disagreement is an obstacle for democracy is tacitly embracing a contradictory conception of democracy: "Democracy without disagreement. That's no democracy at all" (Aikin & Talisse 2020: 40). Thus, insofar as none of the views involved in a disagreement attacks the basic principles of any democracy, or stops recognizing their adversaries as political equals, the distance or whatever parameter deemed relevant in order to conceive polarization in terms of how beliefs and opinions are represented in an ideological spectrum does not strike us as a good explanatory tool to tackle the harmful nature of political polarization.

Moreover, in the next chapter we will see that polarization, understood as changes in belief contents, does not fit so well with certain powerful evidence about some mechanisms and phenomena that promote polarization. We will see that the way in which some of these mechanisms and phenomena work is not by mostly promoting changes in the contents believed, but by increasing confidence in the contents already believed. In that sense, ideological polarization can hardly accommodate some of the evidence about how polarization occurs. Since these phenomena and mechanisms have been extensively studied and there is a lot of evidence confirming their relationship with the increase of polarization, the inability of ideological polarization to accommodate them will be an additional problem for the concept.

Finally, we would like to note that some studies which conceive polarization as ideological polarization have suggested that countries like the United States and the United Kingdom, which appear to be highly polarized given the problems they face, are not actually polarized (Fiorina & Abrams 2008; Fiorina et al. 2008; Fiorina & Levendusky 2006; Fiorina 2017; Levendusky 2009; Wolfe 1998). Thus, it seems that, given the amount of evidence confirming the difficulties that many democracies face, and the intuition that some of these difficulties are related to polarization of the public, if the concept of ideological polarization leads to think that there is no polarization in these countries, then the conclusion should be that this concept is not a good one to measure and understand the current state of polarization.

## 3.2. Trying to pull up: Affective Polarization

Ideological polarization seems to entail cognitive polarization inasmuch as it is about cognitive mental states, i.e., belief-like mental states. As we have seen in section 2.4, the concept of ideological polarization measures polarization by examining, in a way or another, the contents of self-reported beliefs. However, as we advanced in section 2.6, there can be polarization not just about belief-like mental states, but also about desire-like mental states, in particular about feelings, prejudices, and other kinds of evaluations that people from one group make about their political opponents but also about those within their people (Iyengar et al. 2019). This kind of polarization, which we have called conative polarization, is what corresponds to the traditional notion of affective polarization. In the literature about political polarization, affective polarization has been recently emphasized on the mass-level as a different form of polarization from the concepts of ideological polarization (see Gidron et al. 2020: 3).

Let us state once again our view on this matter. We will not deny that affective polarization is about feelings, prejudices and evaluations. In fact, we think that this is the right way to measure at least one of the types of polarization that endangers democracy. It is our contention, though, that it does not follow from this that affective polarization does not rely on beliefs. It only follows that it does not necessarily have to do with belief contents. It may be completely independent of belief contents, but this is not a necessary condition. Our contention, then, will be that affective polarization, or at least some types of affective polarization, indeed might have to do with beliefs, but it has *more* to do with degrees of belief, i.e., with radicalism. In particular, we will contend that affective polarization measures people's attitudes connected with their level of confidence in some core beliefs of their political identity.

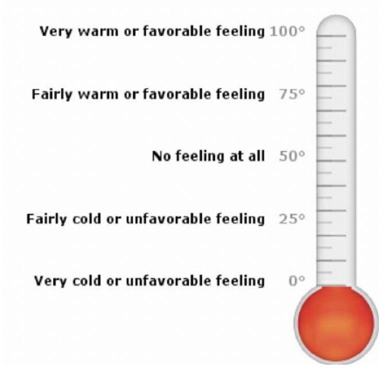
As advanced at the beginning of the chapter, in this dissertation we put under suspicion the widely accepted idea that belief-like and desire-like mental states have a sharply different nature regarding their connection with action. Thus, the difference between ideological polarization and affective polarization, we will argue, is not that the former entails cognitive polarization and the latter conative polarization, but that the former mostly measures, and try to measure, *the mental states people self-ascribe to themselves*, and the latter *the mental states people express to be in*. Of course, this is not to say that the tools to measure ideological polarization always measure the belief contents that people self-report. Someone might express the mental state in which she actually is simply by self-ascribing it, or by self-ascribing a different one. Even, one might express certain affective attitudes, those especially linked to what can be expected from that person, by a belief self-ascription. Our point is simply that the tools employed by affective polarization most frequently target at what people express about their mental states, and not to what they say about how they conceive their mental life to be. But before doing all this and arguing why this is an advantage for the concept of affective polarization that we defend in this dissertation, we must first present affective polarization as it is usually conceived, discuss its limitations, and analyze in more detail the philosophical assumptions it shares with the concept of ideological polarization.

Let's return to the example introduced at the beginning of the chapter. Cayetana's sympathizers exhibited negative attitudes and reported having feelings of hatred and contempt toward Carmena's group, at first glance simply because the latter made some changes in the style of a parade. So, polarization in feelings and attitudes intensified, or turned out to be considerably high. This kind of visceral rejection of anything a person from the "other group" does counts as affective polarization.

Affective polarization, as it is understood, is the tendency to dislike those deemed as opponents in identity terms or "to view opposing partisans negatively and copartisans positively" (Iyengar & Westwood 2015: 691; see also Hetherington et al. 2016; Iyengar et al. 2012; Iyengar et al. 2019; Lelkes 2016; Mason 2013, 2015; Reiljan 2020). Thus, it is defined as the difference between partisans' feelings toward in-group versus out-group (Gidron et al. 2020: 13). As noted, affective polarization is often conceived as based on group affiliation, i.e., affiliation to a social or ideological group. As Iyengar, Good and Lelkes hypothesized, "the mere act of identifying with a political party is sufficient to trigger negative evaluations of the opposition" (Iyengar et al. 2012: 3). These evaluations include considering members of groups deemed as opposite as hypocritical, selfish, and closed-minded (Iyengar et al. 2019). Mason, who calls affective polarization 'social polarization', argues that it involves three different phenomena: implicit biases, emotional reactivity and activism (Mason 2018). As we will see in this section, all these phenomena are measured by different tools.

The basic idea of affective polarization, then, is that we tend to see people who agree with us as one of us, and such identification has affective and behavioral implications, namely, a disposition to like and favor the in-group and to dislike the out-group. In this sense, affective polarization can be understood as the gap between in-group identification and out-group bias (see Harteveld & Wagner manuscript). Affective polarization can describe a country, a set of parties or an individual (Harteveld & Wagner manuscript). At the individual-level, affective polarization has to do with the dislike toward the out-group party and the like toward the in-group. At the group-level, affective polarization has to do with how affectively polarized two political parties are on average. Finally, at the country-level, affective polarization points to how affectively polarized a population is on average toward different out-groups.

Regarding the ways through which affective polarization is measured, we can first distinguish three main tools:*feeling thermometers*, *stereotype tests*, and *feeling linked to situation questionnaires* (see Iyengar et al. 2012: 7). A feeling thermometer is a scale from 0 to 100 through which participants can rank how they feel regarding a particular issue, group, politician, etc. Usually, 0 means "cold", indicating disapproval, and 100 "warm", indicating approval. As to stereotypes take, for example, the Almond and Verba study (Almond & Verba 1963) which asks participants to think about supporters of different political parties, and then rate them by selecting some expressions from a particular set. This set comprises positive and negative expressions, such as “intelligent people”, “interested in national strength and independence”, “selfish people”, “betrayers of freedom”, and “ignorant and misguided”. Finally, questionnaires about feelings linked to certain situations ask questions like “how would you feel if you had a son or daughter who married a republican/democrat (conservative/labor) supporter?” Some options are offered ranging from very unhappy to very happy.


Very warm or favorable feeling 100°
Fairly warm or favorable feeling 75°
No feeling at all 50°
Fairly cold or unfavorable feeling 25°
Very cold or unfavorable feeling 0°

These three ways of measuring affective polarization can be seen as *survey self-reports* inasmuch as they explicitly ask participants to provide information about their feelings and evaluations. Despite these basic and widely used tools to measure affective polarization, there are also other ways to measure it: *implicit bias tests* and *behavioral measures* (see Iyengar et al. 2019 for a review). Implicit bias tests measure the reaction time in associating in-groups and out-groups to positive and negative words. That is to say, it compares the time employed to pairings, for instance, [democrats, good] with [republicans, good] as well as [democrats, bad] with [republicans, bad] (Iyengar & Westwood 2015: 692). This kind of implicit measurement is more difficult to manipulate, but their results are prima facie more valid and less biased than the explicit ones (Iyengar et al. 2019; Iyengar & Westwood 2015), in that they avoid some of the troubles mentioned for self-reports above. The other implicit or indirect measure, the behavioral one, consists in analyzing the participants’ behavior in economic games like the trust game and the dictator game. In particular, they analyze whether participants are willing to endow or withhold financial rewards based on whether the players are in-group or out-group (Iyengar et al. 2019; Iyengar & Westwood 2015).

In sum, affective polarization seems to focus on the feelings that the population has toward people seen as opponents. These feelings are sometimes analyzed using explicit measures, such as self-report surveys, and sometimes using implicit measures of bias and behavior. Hence, the concept of affective polarization seems to be defined as the concept of polarization that has to do with people's feelings, and other goal-oriented mental states, through different types of tools. Does affective polarization, through its different tools, offer an accurate way to measure the existent level of polarization in a country? Before discussing some possible limitations of the concept, in the next section we will try to distinguish different types of affective polarization.

### 3.2.1. Looking into the cabin of a locomotive: Types of affective polarization

In this section we distinguish different possible types of affective polarization. Intuitively, affective polarization is linked both with our feelings towards the members of a different group, and with the confidence that we have on the core beliefs of our political identity. It's common to assume that these go hand in hand, but this is not necessarily so. Our aim here is simply to differentiate some possible situations of affective polarization that will play a role in the dissertation later on, especially in chapter 8. Regarding the argument of this chapter, the main role of the following distinctions is to show that not all interesting varieties of affective polarization share the problems that we have highlighted above and will introduce below for affective polarization, understood as the presence of certain feelings that can be directly reported by the subjects themselves.

The first thing to notice is that not every case of animosity counts as an instance of affective polarization. I might have negative feelings toward New Balance shoes, or toward certain people who use them frequently or think that they are nice and worthy, without necessarily being affectively polarized. There need to be at least two groups of people who are somehow at odds. So this case would not count as one of affective polarization, even if my feelings toward these shoes become increasingly virulent.

The first type of affective polarization that can be distinguished is one where members of a group dislike and hate those who belong to the opposing group simply because they are from that particular opposing group. For example, perceiving supporters of the *Partido Popular* party as spoiled childish people who think that the world belongs to them, or perceiving people who self-identify with the *Unidas Podemos* as rude and crusty-looking would count as an example of this sort of affective polarization. A similar example might be the hooligans of two football teams considered enemies, such as Real Madrid and Barcelona F.C. In these cases we have only animosity toward the other party. We can call *affective polarization with animosity* this type of polarization. Every kind of affective polarization, though, as we will show below, does not require animosity. Perhaps surprisingly, quite politically disruptive forms of affective polarization might be found in the absence of animosity.

A second type of affective polarization is one where members of a group dislike and hate those who belong to the opposing group essentially because they have a high level of confidence in certain beliefs that are central to the identity of their group. An example might be supporters of *Vox*, who have a high level of subjective confidence in some core beliefs of their political identity, such as that men are discriminated against, that other political groups want to break up Spain, or that Muslims want to invade the country (as being Muslim makes you not Spanish), and as a consequence they don't really engage with others' arguments and reasons, becoming impervious to the reasons coming from others. In this case there is not only animosity, but also radicalism. Our initial example also belongs to this kind of affective polarization that encompases hatred towards the other group, but also an inability to engage in a meaningful discussion of the reasons that support the others' position. We can call *affective polarization with animosity and radicalism* to this type of polarization.

A third type of affective polarization is one where members of a group do not dislike or hate those who belong to the opposing group, but simply have a high level of sympathy and support toward people that belong to their own group.

An example might be the Black Lives Matter movement, where for the most part there is neither animosity nor radicalism but still we could talk of affective polarization inasmuch as there is a high level of sympathy and support toward people belonging to the group, and this establishes a clear division between in-group and out-group. We can call this *affective polarization via sympathy*.

A last type of affective polarization is one where there is no animosity but radicalism between two groups somehow at odds. An example might be our disposition to disregard the arguments and the alleged evidence of flat-earthers or advocates of homeopathy, and vice versa, because we have a high level of confidence in our beliefs about the Earth's shape, and homeopathy, and consequently are partially impervious to the reasons of the other part. We can call *affective polarization with radicalism* this type of polarization. As pointed above, when politicians exhibit a personal liking towards members of the opposing party, while at the same time completely disregarding their reasons, they showcase this sort of affective polarization, and can be extraordinarily disruptive to the functioning of certain democratic institutions.

As it can be seen, not all types of affective polarization involve animosity. Moreover, it seems that all forms of radicalism are forms of affective polarization but not all forms of affective polarization are forms of radicalism. It is important to establish these distinctions because, presumably, not all of these types of affective polarization entail the same risks to democracy, when they do, and the ways to measure one type or another, but also the way to intervene to alleviate each of them, should be significantly different.

### 3.2.2. Too low, terrain! Affective polarization and its limits

Having discussed some of the possible types of affective polarization, we now turn to the question of whether the concept of affective polarization, as commonly understood, is problematic.

The first limitation we want to introduce is one pointed out by Fiorina. According to him, it is not clear what is exactly measured by the feeling thermometer, one of the most widely employed tools to measure affective polarization. The objection goes as follows: the reasons why participants may say that they have cold or warm feelings toward a particular person or issue can be very diverse (Fiorina 2017: 59-60; see also Druckman & Levendusky 2019 for a more recent discussion about a similar objection). For example, a citizen can feel cold toward a politician because she thinks that the politician is a bad person but, at the same time, can feel warm toward the same politician due to her foreign policy. Therefore, Fiorina argues, the feelings thermometer does not allow us "to separate the affective from the cognitive" (Fiorina 2017: 60). In other words, we cannot know whether respondents feel what they say they feel due to ideological or moral reasons. And the same can be objected to other measurements of polarization: they do not grant access to the reasons why participants respond in the way they do.

It can be argued that this first objection is not a very powerful one to the extent that, despite the possible uncertainty regarding the underlying reasons, the tools might actually serve to successfully measure people's feelings. In this sense, they would serve to measure affective polarization. However, this objection can be posed in a slightly more powerful version. What does it mean to say that participants report to have cold feelings? Does it mean that they feel hate? Repulsion? Both? And why not just indifference or disgust? How can we know it?

Another concern is related to the connection between the different phenomena measured by the tools. In particular, to the reasons why they are connected, if they are (see Druckman & Levendusky 2019). How can the connection between biases, behaviors and feelings be explained? Why are they all part of affective polarization? A reasonable response might be just to say that they all are affective polarization because, put together, they entail a particular negative view of the other side. But still, why must they be connected? One may object that a person can be biased toward a group of people and feel nothing. And, on the contrary, one may also argue that a person can have cold feelings toward a group of people and, at the same time, behave in a non-biased way. Hence, it seems that the connection demands a particular theory about feelings and mental states. In particular, one that links them with action, and explains the possibility of situations as those mentioned above.

Even if we had such a theory, another related concern may be the following.

It seems that if people feel hate toward the other side, and affective polarization is orthogonal to ideological polarization, then it follows that people can feel hate without any substantive reason, just as a visceral reaction. Thus, the picture this conception of polarization promotes is one according to which people are irrational. Of course, there are arguments in favor of such an explanation. In fact, this is what seems to be assumed by the claim that affective polarization is based on "the tribal nature of intergroup dynamics" (Lauka et al. 2018; see also Iyengar et al. 2012; Lupu 2015; Mason 2013, 2015, 2018), or by the claim that "we believe indications of polarization ought to be rooted in how people feel rather than in how they think or where they stand"(Hetherington & Rudolph 2015: 26). In this line, some recent work explains the current level of polarization that afflicts many contemporary democracies in terms of epistemic vices, intellectual defects (Cassam 2019; Lynch 2019). When some epistemic vices are seen as the causes of political polarization rather than as their outcomes, a certain irrational explanation of polarization is assumed, to a greater or lesser extent.

We acknowledge that stressing the vices that plague our reasoning has certain explanatory advantages, and indeed brings out some of the attitudes which are essential to explaining the current situation (see section 8.3). However, we believe that putting too much emphasis on these issues while explaining polarization, especially as causes of it, carries other risks that are not minor. We leave this question for chapter 8. For the moment, we just want to make the following point. If we pay careful attention to the behavior of an ardent supporter of a view, we can see that she relies on information, maybe not in the best pieces of information, but information after all. In fact, she will presumably share news supporting her ideas, produce discourse and, maybe vehemently, continuously explain why she is right, or why she thinks she is right. The point we are trying to emphasize here is that when someone reports to feel hate towards others it is not necessarily because she is irrational, or because she does not care about truth, but because she thinks she is clearly right, and because indeed she cares about truth.

In fact, it is practically a nonsense, or at best a contradiction, to say that someone believes that p and at the same time doesn't care about the truth of p. The expression "don't care about truth" just can be used with some sense if attributed to others, and the reason is that anyone who holds an idea usually thinks that truth is on their side. To say that someone does not care about truth is to make a negative evaluation, as much as saying that someone is irrational (see Frápolli & Villanueva 2018). As Dorst puts it, "as far as the psychological evidence is concerned, the "other side" is no less rational than you–so if you don't blame your beliefs on irrationality (as you can't), then you shouldn't blame theirs on it either" (Dorst 2020). Hence, it is hard to think that polarization is just about feelings, or even mainly about feelings. Polarized people hold positions and think that they are right. Therefore, there must be a connection with beliefs.[1]

In chapter 8 (section 8.1), we will try to deal with these three objections. The reason is that, as noted earlier, in this dissertation we contend that affective polarization is on the right track to successfully measure the pernicious type of political polarization. The main argument for this, as advanced in the introduction, is that affective polarization tools involve evaluative uses of language, and through that kind of language we express our mind, our affective attitudes.

## 3.3. A picture held us captive: Philosophical assumptions behind the concepts of ideological and affective polarization

It seems that ideological polarization can neither measure people's attitudes in an accurate way nor explain why having different belief contents is necessarily pernicious to democracy (see section 3.1). Affective polarization, on the other hand, seems better positioned to explain why political polarization is pernicious.

[1] Rogowski and Sutherland (2016), and Webster and Abramowitz (2017) have argued that there is a close connection between the rise of affective polarization and the growth of ideological polarization. But this is not the standard explanation of how affective polarization increases. Moreover, their analyses hinge on the idea that extreme ideological positions predict negative evaluations toward the opponents. Our explanation of this fact is that, since to choose an extreme option in a self-report questionnaire about ideology one must overcome all the negative issues it implies, one only chooses an extreme option in these questionnaires when one is very convinced of what one believes, that is, when one has a high credence in the core beliefs of her ideological identity.

In that sense, it is more promising. However, affective polarization, as it is commonly understood, has similar problems concerning the accuracy of their explicit measurements (see section 3.2.2). Moreover, it depicts an image of polarization that, understood just in terms of feelings, seems inadequate, or at best incomplete.

In this section, we will be concerned with the philosophical background behind both concepts. Beyond their problems, the main difference between ideological and affective polarization concepts seems to be that ideological polarization tries to measure what people *believe* by directly asking them about their beliefs –and in that sense it tries to measure cognitive mental states through belief self-attributions, while affective polarization tries to measure what people *feel* mainly by directly asking about their feelings –and in that sense it tries to measure noncognitive mental states mainly through feelings self-reports. In what follows, we make explicit two philosophical assumptions behind both concepts of polarization. In particular, we stress that they assume a *difference in nature between belief-like and desire-like mental states*, on the one hand, and that there is a strong kind of *first-person authority* regarding our own mental states, on the other.

### 3.3.1. I'm doing nothing, just having a belief! Reason vs. the heart

Beliefs, opinions, thoughts, ideas, knowledge, and other similar mental states are traditionally considered radically different from desires, expectations, feelings, wishes, hopes, fears, and other similar mental states. This distinction, as we will see, is very intuitive at first glance and is widely accepted in philosophy. To introduce this mainstream approach, consider the following situation.

Suppose I believe that there is beer in the fridge. In that case, it seems that if I open the fridge and there is no beer there, then my belief is false. On the contrary, if there is beer in the fridge, then my belief is true. Thus, what I believe can be true or false depending on how the world is. Now suppose that instead of believing that there is beer in the fridge, I desire to have a fridge full of beer. In that case, my natural course of action, other things being equal, would be to go to the shop and make sure that I buy enough beer to fill my fridge. Thus, what I desire can be fulfilled by making the world fit my desire. Despite the fact that both beliefs and desires seem to be somehow touched by the world, it is assumed that beliefs are in the business of truth and falsity, while desires can just be fulfilled and do not have truth-conditions: beliefs are about how things in the world are, while desires are about how one would like the world to be.

A traditional approach to mental states, which naturally fits with this way of explaining the nature of mental states, is representationalism. Matthews calls it *the received view* (Matthews 2007: 19). According to this view, belief ascriptions point to mental representations, which represent states of affairs, i.e., particular distributions of objects, and their truth depend on the representation of the state of affairs meeting the way the world is.[2] The picture promoted by this approach looks like a box inside our head in which we store the representational contents of our mental states, and when they are reached by the world they turn out true or fulfilled, depending on the type of mental state.

This distinction between beliefs and desires is often explained in terms of their different "direction of fit" with the world. Take, for instance, the following quote.

> The distinction is in terms of the direction of fit of mental states to the world. Beliefs aim at being true, and their being true is their fitting the world; falsity is a decisive failing in a belief, and false beliefs should be discarded; beliefs should be changed to fit with the world, not vice versa. Desires aim at realization, and their realization is the world fitting with them; the fact that the indicative content of a desire is not realised in the world is not yet a failing in the desire, and not yet any reason to discard the desire; the world, crudely, should be changed to fit with our desires, not vice versa. (Platts 1979: 257)

The point is that beliefs should fit the world, and they should be changed if they do not. Desires, however, should not be changed if they do not fit with the world –it is in fact impossible to desire what we know is actually the case; it is the world what should be changed. In that sense, they have a different direction

[2] This idea has been versioned and refined in various ways (see, for instance, Forro 1987; Harman 1973; Loewer & Rey 1991; Pylyshyn 1984; Sterelny 1990).

of fit: beliefs should fit the world, while the world should be made to fit our desires. Hence, beliefs are about the world, desires express motivations to change the world, to do something. Williamson (Williamson 2002) puts this difference not in terms of truth-aptness vs. satisfaction, because desires might be satisfied and beliefs might be true just by chance, but in terms of knowledge vs. action: "The point of desire is action; the point of belief is knowledge" (Williamson 2002: 1). Beliefs aspire to knowledge, while desires aspires to action. The point is that desire-like mental states, in contrast to belief-like ones, seem to be intrinsically action guiding.

The distinction between cognitive and noncognitive or conative mental states is based on a very similar idea: a mental state is a cognitive one when its content can be true or false, because it can be known. That is to say, beliefs are the attitudes we have when we take a proposition to be the case or regard it as true. Noncognitive mental states, on the contrary, instead of being true or false, are motivations to do something. This way of seeing the difference draws on the Humean idea that beliefs alone are incapable of motivating action, which is known as the Humean theory of motivation (see Smith 1987), and also on the debate regarding moral vocabulary and other uses of language such as aesthetic ones: cognitivists have traditionally maintained that moral claims are truth-apt and express belief-like mental states, while noncognitivists, as the canonical reception of the tradition known as emotivism or classical expressivism (Ayer 2001; Stevenson 1937; Ogden & Richards 1923) pictures them, deny it, and maintain that these claims express the speaker's pro or con attitudes or emotions, that is, their approval or disapproval of something (Ayer 2001: 109).

Thus, a defining feature of noncognitive mental states is that they have an extra motivational component compared to cognitive ones. It seems that when I say I desire, want, fear, hate, etc., that p, the content of my mental state is not true or false, but it expresses my motivations, what I am willing to do, the courses of action that can be expected from me, my emotions, my evaluations. This conception fits very naturally with the idea, widely repeated in the literature about affective polarization, that animosity plays a key role in motivating partisan behavior (see, for instance, Huddy et al. 2015).

It is important to note that the distinction between cognitivism and noncognitivism can be conceived as a semantic distinction or as a psychological distinction (see O'Leary-Hawthorue & Price 1996: 276; Pinedo 2020). From a semantic point of view, a claim is cognitive when it has truth-conditions. From a psychological point of view, a mental state is cognitive when it is belief-like. We have deliberately introduced the cognitive vs. non-cognitive distinction without distinguishing these two approaches because they are often assumed together (see, for instance, De Mesel 2019). In fact, cognitivism, as it is traditionally[3] uunderstood, entails other independent theses that are normally mixed within it, that is, *representationalism*, *descriptivism*, *realism*, and a *correspondence theory of truth*. The mix of both distinct levels of analysis is based on the idea that through the descriptive use of language, aimed at informing about the world, we express belief-like mental states, while in using the language in an evaluative way we express desire-like states of mind.

We will briefly discuss some of these theses in chapters 6 and 7, and argue that there is no sharp division between belief-like and desire-like mental states regarding their action-guidance and truth-aptness: many belief-like mental states are linked to certain courses of action, and their truth or falsity do not always depend on how the world is. On the other hand, many desire-like mental states are not especially significant regarding their connection with action, and many of the claims that express a desire-like mental state can be declared true or false. Hence, although we acknowledge that there are differences between beliefs and desires, we will argue that the allegedly different nature regarding their truth-aptness, if any, as well as their connection with action, is actually in their *contents* or the *claims that express them*, not in the *type* of mental state they are (see chapters 5, 6 and 7). This will allow us to reassess the concept of affective polarization, in a way that successfully meets our desiderata.

[3] There are more contemporary cognitivist positions that do not assume all the associated theses, such as certain cognitivist expressivism.

### 3.3.2. Because I say so: First-person authority

Another philosophical assumption of both ideological and affective concepts of polarization is that they take participants' self-reports to be a reliable source for knowing their mental states, be them beliefs or others. In that sense, both concepts presuppose that people somehow have *authority* to determine the state of mind they are in. Take the following Finkelstein's quote to illustrate the kind of authority that mental state self-ascriptions allegedly exhibit:

> If you want to know what I think, feel, imagine, or intend, I am a good person –indeed, usually the best person– to ask. It is sometimes said that I enjoy a kind of authority when I talk about what, broadly speaking, might be called my own states of mind –when I say, e.g., "My head hurts", "I was worried about you", or "I intend to arrive early". When people don't accept my mental state self-ascriptions at face value, it is generally because they take me to be insincere rather than mistaken. (Finkelstein 2003: 9)

The special authority that mental self-ascriptions seem to exhibit is well-known and discussed in philosophy (see, for example, Bar-On 2004; Barz 2018; Borgoni 2018a; Coliva 2016; Davidson 1984; Villanueva 2014; Wright 1998). In particular, the thesis that mental state self-ascriptions have a presumption of truth as long as they are sincerely made is known as *first-person authority*, one of the features explaining the alleged special character of self-knowledge. Note that in this case, what is deemed as true is not the content of the mental state, but the self-ascription itself, i.e., whether a person has a particular mental state or not, and not whether her belief is true or false. Thus, according to the authority thesis, the speaker's sincerity in self-ascribing a mental state guarantees that it is true that the speaker is in the mental state she says to be in. The view behind this thesis is quite intuitive and, also, politically powerful: if you want to know what someone believes or feels, the best way is to directly ask her and trust her answer (Falvey 2000: 69). Let's consider an example to see how intuitive is the idea that subjects exhibit some kind of authority regarding her own mental states.

Take a belief self-ascription such as my utterance of the sentence "I believe that there is beer in the fridge". It seems intuitive to think that if I sincerely utter that sentence, then it is true that I believe that there is beer in the fridge. On the contrary, if I attribute a state of mind to someone by sincerely uttering the sentence "My colleague believes that there is beer in the fridge", it seems that my sincerity does not guarantee, in the same sense, the truth of the proposition expressed. It is precisely my colleague who should speak to the fact that she believes or not. Thus, it appears that one is in a better position than anyone else to talk about one's own mental states.

First-person authority is one of the features usually deemed as essential for self-knowledge,[4] which is understood in contrast to our knowledge of the external world, and also to our knowledge in other realms. In that sense, the received explanation of the authority feature of first-person is epistemic (Hacker 2005). The idea is that there is an asymmetry between the way we gain knowledge about our own mental states, and the way we know about the world and other minds (see, for instance, Byrne 2018). That's the reason why it appears that third-person mental-state ascriptions, in contrast to first-person ascriptions, are not authoritative. However, note that these are two different senses of authority. One simply has to do with the *sincerity* of the speaker. The other has to do with the privileged way through which agents *know* about their own mental life. In other words: the authority feature can be epistemic or not, i.e., it can rely on a special method of access to one's own mental states or not. Henceforth, when we talk about author-

[4] Three other alleged features of self-knowledge are that it is *especially secure*, that is acquired by a *special method*, and that is *transparent*. Cartesianism is the paradigmatic position defending the authoritative character of first-person self-ascriptions that also holds transparency, the special security and the special method of self-knowledge. According to this view, the speaker's word is authoritative about her mental states because she has a *direct*, *transparent* and *privileged access* to that part of her *inner life*, in particular through introspection. This view, as it can be deduced, is a representationalist one: it contends that we have internal representations to which we have privileged access.

Regarding 'transparency', the sense of which we refer to is the idea that it makes no sense to wonder whether the speaker knows that is in the state of mind she says to be in (Wright 1998: 15). This sense is radically different from Evans' concept of transparency, which is transparency toward the world (Evans 1982), an anti-Cartesianism type of transparency.

ity, we are talking about the former non-epistemic kind. How can someone have authority without having epistemic authority? As Bar-On argued, for instance, you might be simply trained in such a way that your self-attributions only make sense when they actually express the state of mind you are in (Bar-On 2019) (see chapter 5).

At least two different types of first-person authority can be distinguished in terms of their degree of restrictiveness (see Falvey 2000: 70; Villanueva 2014: 52-53). On the one hand, according to the strong version, the speaker's sincerity always entails the truth of her mental self-ascription. That is to say, the speaker's word is all that is relevant, in every case, in order to determine the truth or falsity of a mental self-ascription. This kind of authority is a hard thesis to maintain, given the existence of phenomena like self-deception (Mele 2001). On the other hand, we have a weak version of the authority thesis. According to this second version, in spite of the fact that external evidence can sometimes show that a particular sincere self-ascription is not true, there is an assumption in place that every time a speaker self-ascribes a mental state, its sincerity guarantees the truth of her self-ascription. This presumptive authority is a more widely accepted view.

Since the main tools to measure ideological and affective polarization are direct self-report questionnaires, both concepts share the presumption that participants' sincerity in self-ascribing a mental state guarantees that they actually are in the mental state they say to be in, at least in its weak version, and regardless of whether it is a belief or an affective mental state. Hence, this is one of the philosophical assumptions that are taken for granted when these concepts of polarization are put to the test in empirical studies.

Before ending this section, we want to make it clear that, even though the idea that the speaker's sincerity guarantees the truth of a mental self-ascription is challenged all throughout this dissertation, we don't argue against the idea that we must trust people's claims regarding their own mental states in our daily contexts (see chapter 5). We review some empirical evidence and theoretical arguments that, taken together, put under suspicion the presumption that the speaker's sincerity guarantees the truth of her self-ascriptions, even in the weak sense. And since political polarization is a very dangerous phenomenon that has to be detected as early as possible, this outcome must be taken into consideration regarding how to measure polarization. But note that this is compatible with the political stance of taking as true what people say about their mental life in other contexts, and with the psychological benefits of trusting others (Yamagishi 2001; Yamagishi et al. 2002).

## 3.4. Shaking the assumptions: A first attempt

In this section we will make a first attempt at challenging the plausibility of the philosophical assumptions behind ideological and affective polarization concepts: 1) there is a difference in nature between belief-like and desire-like mental states, and that 2) there is a strong kind of first-person authority regarding our own mental states. Let's discuss the authority thesis first. In particular, let's discuss some specific situations keeping in mind the idea that the speaker's sincerity guarantees the truth of her self-ascription.

### 3.4.1. Even if I say so

There seem to be situations where we sincerely make different claims about our mental life that are not easily compatible with each other. In particular, sometimes we say that we believe that p but we behave, verbally and non-verbally, in a way that hinders the plausibility of saying that what we say we believe is what we actually believe. For instance, according to Ellis and Stimson, when Americans are asked about their particular policy preferences they tend to choose liberal positions instead of conservative ones. However, when they are asked about their ideological identity, they tend to self-identify as conservatives (Ellis & Stimson 2012; see also Mason 2018). Since identifying as liberal or conservative is presumably closely linked with having certain ideological preferences on some specific issues, it is only to be expected that if one self-identifies as liberal, then one will mostly choose liberal positions related to those specific issues. Then, if someone sincerely chooses liberal positions and sincerely identifies as conservative, what does she really believe?

In order to explain this, Ellis and Stimson distinguish between "symbolic ideology" and "operational ideology". The first one is just a matter of identity, while the latter refers to a set of policy positions. Mason emphasizes this difference by saying that it is important to recognize that ideological identity is not synonymous with political preferences (Mason 2018). We agree that some of our ideological responses do not convey our policy preferences; we can convey another kind of information with them. In particular, we can express the level of confidence in some core and salient beliefs of a particular ideological identity. But this is not what the participants are told before giving their answers, and therefore they are not aware of that. In that sense, this common situation suggests at least that it is not always clear that what we say we believe is what we actually believe.

Recently, philosopher Kate Manne has brought domestic inequality to our attention to highlight one of the unjust tenets conforming our misogynistic environment: men do far less than their female partners regarding housework (Manne 2020: 166-188). This situation is well-known and widely supported by several empirical studies (Yavorsky et al. 2015). What is perhaps less known is our particular perception of the situation. According to the results of a study conducted in eight Western countries, 46 percent of male partners reported being coequal parents, while only 32 percent of their female partners agreed with their assessment (Sha 2017). It is not hard to imagine that, although many of the male respondents of this study were sincerely self-reporting what they believe about their situation, their female partners would disagree not only on the truth of the content of their beliefs, but even on whether they really believe what they say they believe.

Another phenomenon, known as the illusion of explanatory depth, goes in the same direction of putting under suspicion the authority thesis. The illusion of explanatory depth, named by Rozenblit and Keil, has to do with the perception of knowing the world with far greater detail and depth than we actually do (Rozenblit & Keil 2002). Through a large number of experiments, people were asked to rate their knowledge of a device, a mental illness, an economic issue, and other similar issues. Then they were asked to try to explain it, and finally rate again their knowledge of the topic they were asked about at the beginning. After being confronted with their inability to explain it in a detailed way, they rate lower their knowledge of it (Fisher et al. 2015; Lynch 2019). Of course, this path can just be seen as a response to changes in the level of explanation expected. However, since this phenomenon happens very frequently, we can fairly wonder whether people believe what they say they believe, at least when they are asked about complex issues. In other words: if someone chooses a specific option as a representation of what she believes about, for example, the economic policies that must be taken in a country, and yet it turns out that she cannot explain what those policies consist in, then does she really believe that those policies should be taken? This path has been replicated on responses about political issues (see Fernbach et al. 2013; Vitriol & Marsh 2018). Hence, the speaker's sincerity does not seem enough to know people's minds, at least in these particular cases.

There is more empirical work casting doubts on the authority in self-ascribing a mental state. Several studies have shown that people have systematic blindspots regarding their own mental states (Nisbett & Wilson 1977; Wilson 2002; Zajonc 2001). Nisbett and Wilson, for instance, conducted some studies whose results show that participants reported that they prefer a particular product over its competitors because of its apparent quality, when in fact it was the spacial position of the product that influenced their choices (Nisbett & Wilson 1977). Thus, if someone says she believes that a specific product is better than the others because of its quality and in fact it was its position that influenced her choice, does she really believe that the specific product is better than others because of its quality? That's the reason why confabulating poses a threat to self-knowledge (Carruthers 2011), and derivatively to authority.[5] . In a quite similar line, research on punitive policies (Carlsmith et al. 2002) and racist behavior (Dovidio et al. 2002) revealed that we say we endorse some standards which our behavior does not reflect. It seems that empirical psychology pushes toward the direction that we are systematically

[5] Some authors have recently argued that confabulation does not necessarily undermine first person authority if it is conceived as a capacity for self-regulation, i.e., a capacity "to bridge the gap between our sayings and doings by aligning our actions with our avowed self-ascriptions and vice-versa" (De Bruin & Strijbos 2020: 152). However, with the definition proposed they assume that there is actually a gap between what we say and how we behave and, to our purpose, it is enough to think that confabulation put pressure on the presumption of authority, even we had the capacity to bridge the gap (which is not very clear to us in a large number of cases).

unreliable in grasping our own mental states (Carruthers 2011; Schwitzgebel 2008, 2011a; see also Srinivasan 2015).

Concerning intentions, it is often assumed that doing something intentionally requires knowing (see, for instance, Marcus 2019) or being aware that one is doing it; there seems to be a strong conceptual connection between both things. In fact, in our daily exchanges, saying that one did not know or was not aware of what one was doing is a way of saying that one did not have the intention of doing it. However, a recent study has shown that, at least in some cases, one can intentionally perform an action despite neither being aware nor knowing that one is doing it[6] (Vekony et al. 2020). Thus, in these cases, the sincere self-ascription that one was not aware of or did not know what she was doing, which seems equivalent to say that one had not the intention, does not guarantee that in fact one did it without the particular intention.

Regarding feelings and other allegedly motivational states, things do not seem really different (see Wilson 2002). There are lots of situations in which we are not able to identify what is exactly that we are feeling. And, when we move from one feeling to another, we are often unable to recognize the change. A good number of studies have replicated this effect: although participants act as if they had a certain feeling, usually induced, they do not report the existence of such a feeling (Schachter & Wheeler 1962; Schwitzgebel 2008, 2011a).

For our part, we have conducted an empirical study on the influence of different contextual factors in assessing the same statement as offensive or simply informative, in which we have also observed a particular mismatch between participants' abstract and concrete judgments (Almagro et al. forthcoming). In par-

[6] In the so-called Dreyfus-McDowell debate on the rationality of unreflective action, John McDowell (McDowell 2008: 368-369) makes the claim that the minimum condition to consider certain action as rational is wonder if the question "Why did you do that?" makes sense. If this question is intelligible, then the action can be seen as rational, no matter whether it was unconsciously performed or not (see also Pinedo 2018). For instance, a person that catches a frisbee coming toward her, performs a rational action (McDowell 2008: 368-369). Following McDowell, then, it makes sense to think that a person can perform an action unconsciously but intended, especially in practical and expert knowledge contexts, where a subject develops ways of doing things that she is not aware of because she has not the need to stop and think about what to do before doing it.

ticular, we designed a first study with a set of 8 vignettes in which a speaker utters a sentence, and manipulated three factors: speaker membership (the speaker belongs or does not belong to the group of people she is talking about), speaker intention (the speaker has the intention to offend or not), and outcome (the public is offended or not). All participants viewed 8 vignettes telling a short story in which a speaker claims something, and each participant was exposed to a randomized combination of factors in each vignette. The variation of factors was designed to be relatively imperceptible: they are simply presented along with other details in the story. Then, we asked each participant to rate from 1 (strongly disagree) to 7 (strongly agree) the following claims after reading each scenario: "The speaker simply offers information"; "The speaker is being offensive". We added an additional question about the acceptability of the claim with two available options: "The speaker shouldn't say that kind of thing" / "I don't see any problem in the speaker saying that kind of thing". At the end of the experiment, we included a general final question in which we explicitly asked about the influence of each of these three factors when considering the same statement as offensive or not. Participants had to rate from 1 (little influence) to 7 (much influence) speaker membership, speaker intention, and the harm felt by the audience. Our results showed that for participants, speaker membership played a more important role than the intention and the outcome factors in considering the same statement as offensive and prohibited. However, when explicitly asked, participants rated the speaker's intention factor as the most relevant factor, and the speaker status as the least important factor. This result can be understood as participants *saying* that they believe the intention of the speaker is the most important factor to consider a statement as offensive, while nevertheless *showing* through their answers that, on the contrary, they take the speaker's status to be the most relevant factor. We believe this is the right way to interpret our result because in the concrete responses of each scenario, participants are intuitively responding, i.e., they are evaluating concrete situations, which is closer to measurements of bias and behavior; while in the general final question, participants have to think in abstract terms about what they believe, as they do when responding about their position on political issues, and this kind of reasoning is more prone to error (Iyengar et al.

2019; Iyengar & Westwood 2015).

Moreover, in the philosophical discussion on the second person perspective in knowing our minds (Bilgrami 2006; Darwall 2006; Ferrer 2014; Gomila 2002; Pinedo 2004; Velleman 2009), some approaches have emphasized the authoritative role of the second perspective, sometimes even in a better position than oneself to know one's mind. Thus, it seems that the difference between what we *sincerely say* that we believe, and what we *express* that we believe is more persuasive than it would appear. Consequently, a strong take on the authority thesis must be placed under suspicion regarding the measurement of polarization.

In spite of the literature reviewed in this section, it is worth noting that it is not hard to imagine situations where, from a very intuitive point of view, the mental states we sincerely self-report are not the mental states in which we are. If part, or all, of the literature reviewed here has not convinced you, perhaps you will be a little more convinced by considering the following possible situations. Someone, for example, can sincerely self-report that she is angry, but it could be the case that she is just hungry. Or someone can sincerely self-report that she believes that taxes should be lowered, but it could be the case that she is not ready to act as someone who believed such a thing. Or even it could be the case that through such a belief self-ascription, the speaker is just expressing her adhesion to a certain political ideology, her practical attitudes, instead of providing uncolored information about their minds. These situations are not unconceivable, and they do not seem to be bizarre or marginal situations. Srinivasan appeals to something similar in her discussion about Cartesianism. Let's end this section with a quote from her:

> For my own part I think the most powerful reason to embrace Anti-Cartesianism is (not unironically) introspective. I sometimes find myself uncertain, even after careful consideration, about my own phenomenology: whether I'm angry or merely annoyed, whether I'm desirous or indifferent, whether I believe or am agnostic. Of course, the uncertainty at issue here is not an uncertainty about whether my phenomenology is thus: I'm always in a position to know that I'm

> feeling just this, the way I'm feeling. Instead, the uncertainty lies in the categorisation of my phenomenal experience under the appropriate concepts: anger, annoyance, desire, indifference, belief, agnosticism. My own introspective feelings of uncertainty deepen when we move to conditions of particular philosophical interest, such as my having a credence x in p or p's having probability x on my evidence. For these conditions, I very often feel that no amount of assiduous introspection will reveal whether they obtain. (Srinivasan 2015: 275)

### 3.4.2. I'm doing something, I have a belief!

What about the distinction in nature between belief-like and desire-like mental states? Consider the sentence "the government should go to prison for the deaths caused by the COVID-19". Note that by uttering such a sentence, one is likely to be expressing her desire that the government enters into prison. This is so because one is using language in an evaluative way (see chapter 7). The speaker is talking about how the world should be, and not about how it actually is. That is to say, the speaker is expressing what can be expected from her, maybe joining a demonstration with a golf club and a fancy cerise saucepan in order to achieve a change in the world and satisfy her desires. However, note that by uttering p, we are entailing that we believe that p. In other words, I cannot assert "the government should be put into prison for the deaths caused by the COVID-19" and after that say "I don't believe that the government should be put into prison for the deaths caused by the COVID-19" without triggering a contradiction; if I assert that p, then I endorse that I believe that p.

Of course, as we have seen, it may be the case that I assert that p, and I do not really believe that p. But if I assert p, I cannot say that I don't believe that p without triggering a contradiction, because there is a close link between asserting p and acquiring the commitment that I believe that p is the case, even if it is later discovered that one does not believe p despite asserting it (see chapter 6). Similarly, by endorsing the sentence "The speaker shouldn't say that kind of thing", participants in our experiment entail that they believe that the speaker shouldn't say that kind of thing. However, the claim does not seem to be about how the world is, but about how it should be. Hence, our beliefs are not always about the world, sometimes exhibit a motivational component.

Moreover, note also that we can discuss the truth or falsity of a normative or evaluative claim, that this domain is knowable, and that we can disagree about it. We can maintain that it is false that the government should pay for the deaths by going to prison, or that the speaker should not say this or that kind of thing. Nevertheless, the kind of disagreement in which we enter is different from the one that we enter when disagreeing about whether there is beer in the fridge; the contending parts do not necessarily agree on how to resolve the disagreement, and the disagreement is not straightforwardly factual (Field 2009; see also Osorio & Villanueva 2019). This matter will be discussed in detail in chapter 7. But the thing, at least for the moment, is that many claims, which are not about how the world is, but about how it should be and what we are willing to do, and therefore express desire-like mental states, may, however, be true or false, and be knowable. In addition, by making such claims we also entail that we have such beliefs. Hence, it seems difficult to maintain that sharp distinction between belief-like and desire-like mental states introduced in section 3.3.1.

To put it another way, it seems that the things that we believe, desire, hope, etc., *sometimes* reflect our worldview, our mind, and they are closely linked to certain patterns of action, i.e., with the norms that govern our social practices, our form of life. Our attitudes have to do with how we behave and how we are willing to behave, and this is normatively constituted. From psychology, attitudes are conceived in general in a more motivational fashion: as our evaluations of a person, an idea, an object, or a state of affairs, in a positive or negative way. According to a classic way of putting it, an "attitude is a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (Eagly & Chaiken 1993: 1; see also Albarracín et al. 2005: 4). However, instead of separating between types of mental states by its nature, it is assumed by this perspective that an attitude can be composed by three different components: affective, behavioral and cognitive ones. For example, you may hold a positive attitude toward the changes introduced by Carmena and her team in the Wise men parade. This attitude may result in positive feelings toward those changes (e.g., you say that these changes make you feel good), may be reflected in your behavior (e.g., you attend the parade and promote it), and may also be reflected in your thoughts (e.g., you believe that the parade is really good).

Note that the link with action of an attitude, then, is mostly determined by its content rather than by the type of mental state. For instance, the conceptual commitments that someone expresses by saying that she believes that marriage is the central mark of a dignified life are much more linked to action than those expressed by saying that she believes that the table is brown, especially when in fact the table is brown. Likewise, the conceptual commitments that someone acquires by saying that he wants to get married more than anything else have more links with action than those expressed by saying that he wants peace in the world or to get a brown table for the sitting room. Hence, even if it can be argued that beliefs and desires point to commitments of different kinds, the specific commitments and their links to action mostly depend on what is believed or desired: some propositions have more links with action than others, and some of the first ones are not shared by most people, and maybe because they express particular information about the mind of those who believe or desire them. Of course, there is a distinction in play here. We can distinguish between our *evaluations* of something and our *reports* of something and, if you will, hold that the attitudes we express are different by virtue of their degree of action-guidance. However, this is not a distinction between belief-like and desire-like mental states, but one between descriptive and evaluative information, which is orthogonal regarding the belief / desire distinction. We will come back later to these ideas (chapters 5, 6 and 7).

## 3.5. Pursuing a pathway: Some desiderata for a suitable concept of polarization

Taking into consideration our discussion so far, we propose five adequacy conditions that a notion of political polarization needs to meet in order to be a suitable one, to be operational. The first desideratum we propose has to do with the causes of polarization. In particular, we contend that a suitable concept of polarization must accommodate the evidence regarding how we polarize, that is, it must accommodate the psychological, social and linguistic mechanisms fostering polarization, which will be reviewed in chapter 4, as well as taking into account other kind of evidence we have available related to the issue.

Second, an appropriate notion of polarization must be able to explain the negative effects that polarization poses to democracy, that is, why the raise of polarization weakens respect for democratic norms, corrodes basic legislative processes, fuels public disaffection, increases anger, violence, intolerance and discrimination toward the other side, and diminishes societal trust (Carothers & O'Donohue 2019: 1-2; see also section 2.5).

Third, the pursued concept of polarization should not explain the increased polarization in terms of irrationality or lack of interest in truth. This line of explanation we take to be incomplete at best. On the contrary, it must be able to explain why it is at times rational to become polarized, acknowledging that it is precisely the fact that people care about truth, and that people give reasons in support of their own beliefs, that leads them to polarize; polarization is not just a matter of irrationally hating the other part.

Fourth, an operational notion of polarization must be able to accurately measure people's attitudes. In that sense, the concept of polarization must accommodate the disanalogy between self-ascribing a mental state and expressing a mental state. As we have suggested, people's sincerity is not enough to guarantee that they are in the mental state that they say they are. Hence, indirect ways to measure polarization should be used, ways that involve *the expression of* mental states rather than their self-ascription.

Finally, an adequate notion of polarization must allow us to intervene in a process of polarization as soon as possible, given the special difficulty of depolarizing (see Levendusky 2018; Sunstein 2017) a highly polarized scenario, and given the serious consequences of a high level of polarization (see section 2.6). Thus, the desiderata that we propose a concept of polarization must meet can be summarized as follows.

> **EVIDENCE**: A suitable notion of polarization must be consistent with the best available evidence.
>
> **DANGEROUSNESS**: A suitable notion of polarization must be consistent with the pernicious effects for democracy of a high level of polarization.
>
> **RATIONALITY**: suitable notion of polarization must neither blame people nor account for the issue in terms of irrationality.
>
> **DISANALOGY**: A suitable notion of polarization must accommodate the disanalogy between self-ascribing a mental state and expressing a mental state in order to accurately measure it.
>
> **INTERVENTION**: A suitable notion of polarization should allow us to develop mechanisms to intervene as soon as possible.

In the next chapter, we will see that extremism, or any other mode of understanding the division that characterizes political polarization in terms of belief contents, cannot accommodate well the evidence on how we polarize. In this chapter, we have seen that understanding polarization in terms of belief contents cannot easily explain why polarization is pernicious for democracy (section 3.1). In that sense, ideological polarization seems at first sight incompatible with EVIDENCE and DANGEROUSNESS. Moreover, since ideological polarization measures people's belief through opinion self-report questionnaires assuming that they report what they actually believe, it is not clear how it can fulfill the DISANALOGY condition. Hence, ideological polarization does not seem a suitable concept of polarization inasmuch as it cannot satisfy some, if not all, of the proposed desiderata.

Some authors have argued that the process of discussing with likeminded people is what leads us to be affectively polarized (Aikin & Talisse 2020: 38-39), which renders especially dangerous the rupture that polarization consists in, and that it is in people's feelings about their opponents that polarization manifests (Hetherington & Rudolph 2015: 28), specifically because the increased dislike of the opposite side expresses the decrease in trust in government when the other side is in power (Hetherington & Rudolph 2015: 32). Other authors have argued that it is social sorting and the rise of mega-identities what leads us to be affectively polarized (Mason 2018). Thus, these authors say that it is the mechanism of polarization behind discussing with likeminded people, together with group identity, what raises affective polarization. We agree with them. However, we think that these two mechanisms are just two ways of increasing the confidence in prior beliefs, and that increasing the level of confidence in some core beliefs of our ideological identity is what sometimes leads us to dislike our opponents. But, as we will see in the next chapter, there are also other different mechanisms promoting the increase of confidence in our prior beliefs. Once again, disliking the opponents, and evaluating them in negative terms, could be just a way of expressing our level of confidence in some core beliefs. But then, the pernicious thing is the level of confidence in certain beliefs, from which feelings toward opposite people might be just a symptom. Besides, all the available and relevant evidence concerning the rise of polarization must be approached together: this is crucial to favor an irrational or rational story for polarization processes (see chapter 4).

In this line, we will argue that affective polarization, or some types of affective polarization, measures radicalism and, therefore, is compatible with EVIDENCE and DANGEROUSNESS. We will contend that by asking about feelings and evaluations, affective polarization sometimes measures not only the self-reported feelings and evaluations, but the level of confidence people express to have in certain core beliefs of their ideological identity. In that sense, it also satisfies RATIONALITY and DISANALOGY (see chapter 8). Finally, we will defend that to the extent that affective polarization can accommodate the phenomena of a diverse nature causing polarization, it enables us to indirectly intervene inasmuch as to fight against these phenomena is fighting against polarization. Crucially, we will argue that our notion of polarization enables us to detect polarization soon, that is, before it is too high. Since depolarization is a hard task to achieve especially when the level of polarization is very high, our notion will enable us to intervene. Thus, affective polarization, understood the way we propose, also satisfies INTERVENTION.

## 3.6. Conclusion

This chapter has been devoted to discussing the adequacy of the concepts of ideological polarization and affective polarization in order to measure polarization. In particular, after presenting some acknowledged difficulties of each concept and making explicit the philosophical background of them, we reviewed some evidence against the two philosophical assumptions mentioned. Our aim was to show that if both concepts are understood as they commonly are, then neither of them seems to be adequate to accurately measure what they try to measure. Finally, we proposed five adequacy conditions for a suitable concept of political polarization.

Our proposal is to show that the actual difference between ideological and affective concepts is not that each one measures mental states of a different nature, but that usually they measure different things. In particular, we contend that, while ideological polarization mostly tries to measure the beliefs people *self-report* to be in, affective polarization mostly measures the degree of belief people *express* to have in certain beliefs. Conceived in this way, affective polarization can fulfill the proposed adequacy conditions and, we will argue, it is on the right track to measure polarization in an accurate way. To differentiate our interpretation of affective polarization from the standard interpretation of it, we will call it *polarization in attitudes* (see chapter 8).

To argue this, we will discuss, in the following chapters, the available evidence about how we get polarized, as well as the philosophical assumptions mentioned. We will devote the next chapter to review part of the best available evidence concerning how we get polarized, and will discuss it keeping in mind the desiderata of EVIDENCE and RATIONALITY. Then, we will dismiss the descriptivist approach to mental state ascriptions (chapter 5) and adopt a nondescriptivist approach to them, according to which having a mental state is having a set of conceptual commitments closely linked to certain courses of action (chapter 6). Thus, in order to measure people's mental states, we have to track their commitments and the ways they behave. After that, we will adopt an expressivist approach to evaluative language, according to which through the evaluative use of language we express our attitudes, our commitments especially linked to certain courses of action (chapter 7). In that sense, we will hold, affective polarization measures the commitments to action people express to have, because the tools employed to measure affective polarization involve evaluative uses of language. The commitments expressed by the evaluative use of language, and the commitments in which mental states consist in, are the same: our approach to mental states and to the evaluative use of language are two faces of the same coin.

# Chapter 4

# How We Get Polarized: Mechanisms Promoting Polarization

What leads someone to demonstrate against the alleged existence of a criminal organization dedicated to falsely accusing teachers for financial gain in Ceuta, triggered by a case of which there is not a single piece of information that could lead to suspecting about anything? How does a person decide to demonstrate, carrying a golf club and a saucepan, or in a Ferrari, and yelling that the Spanish government is responsible for the deaths caused by the COVID19 pandemic, while not wearing a mask and not respecting social distance? What leads someone to celebrate the police brutally charging Catalonian people who just were voting and expressing their opinion regarding the future of the territory they live? What leads someone to assault the Capitol of the United States after that the political party that he identifies with lost the election? How do we become so polarized?

In chapter 2, we have showed that polarization of beliefs can occur in terms of extremism or radicalism, that is, in terms of belief content or degree of belief. Furthermore, we have said that the concept of ideological polarization is a broad concept of extremism, that includes different senses of polarization related to belief contents. In chapter 3, we have introduced the concept of affective polarization, as opposed to that of ideological polarization, and have claimed that at least some types of affective polarization have to do with radicalism, i.e., with having a certain degree of confidence in certain beliefs deemed central to a particular political ideology. Moreover, we have proposed a set of desiderata that a concept of political polarization should meet in order to be an adequate one. One of these desiderata requires the concept to being able to accommodate our best available evidence regarding how polarization increases.

We devote this chapter to review and discuss part of the most important evidence regarding political polarization. First, we dwell on two prominent phenomena related to political polarization, that is, *likeminded deliberation* (section 4.1) and *mixed evidence disagreement* (section 4.2), and examine what type of understanding of political polarization these mechanisms contribute for. In the rest of the chapter we consider some other main mechanisms, of a diverse nature, that are somehow involved in the increase of polarization. These mechanisms are divided into three broad categories: psychological mechanisms (section 4.3), social mechanisms (section 4.4), and linguistic mechanisms (section 4.5). In particular, we pay special attention to analyse whether all these mechanisms and evidence are more easily accountable in terms of extremism or in terms or radicalism. Finally, we briefly argue that if the reviewed evidence is approached together, then the rational story of polarization is sounder than the irrational view –according to which, the rise of polarization is the result of our irrational and biased way of processing information, which is important to the desideratum of RATIONALITY (section 4.6).

## 4.1. Belief polarization: Likeminded deliberation

We started chapter 2 by introducing the case of Dereca and Izquerri, who got polarized after discussing with likeminded people about Spanish government's handling with the coronavirus pandemic. We said that deliberation within a group of likeminded people leads to polarize. In this section we discuss it further. For convenience, henceforth we will call this phenomenon *likeminded deliberation.*

The most widely accepted approaches to likeminded deliberation are *social* *comparison theory* and *persuasive arguments theory*, while two other theories have also received considerable empirical support: *social identity theory* and *corroboration theory*. Since these four approaches to the phenomenon are compatible with each other and are supported by solid evidence (see Isenberg 1986 for a review), they all should be deemed as pointing to different mechanisms that play a role in the increase of polarization. We will briefly present all of them and keep them in mind for further discussion about the suitable understanding of political polarization.

The main tenet of social comparison theory (Festinger 1954) is that people want to be socially accepted and perceived favorably by other group members. People tend to appraise their abilities and opinions by comparison with those abilities and opinions preferred by the group that they identify with. Hence, if in a group X the preferred opinion is p, which can be known after group discussion, then individuals often move their judgments in that direction in order to preserve their social image. In other words, they will present themselves as holding more *extreme* views than those individually held, in the direction of the opinions maintained by the group, just to be accepted and be seen with good eyes by the other members (see Broncano-Berrocal & Carter 2021; Sieber & Ziegler 2019; Sunstein 2002, 2017). According to this theory, likeminded deliberation occurs because of our need to preserve our social reputation.

The second well-supported approach to likeminded deliberation, i.e., persuasive arguments theory (Burnstein & Vinokur 1977), is, in a sense, more rational than the previous one: it emphasizes the role of arguments in a deliberative process. The key idea of this theory is that in a collective discussion on two competing alternatives, each member is almost always exposed to new information in the direction of her prior position, and this reinforces her confidence in that position. Being exposed to new arguments supporting a preexisting opinion is quite persuasive. Hence, if in a group X there is a particular opinion tendency, then each member will be exposed, during deliberation, to new information and arguments in support of the dominant opinion. As a consequence, the prior dominant position will be reinforced (Brown 1985; Burnstein & Vinokur 1977; Vinokur & Burnstein 1978; see also Broncano-Berrocal & Carter 2021). The central point of this theory is that when we discuss some issues in a group of likeminded people, there is a limited argument pool pressing just in a particular direction (Sunstein 2017). In other words, it increases the size and density of the pool of arguments that one is exposed to, which fosters polarization. When we are exposed to a limited and biased set of arguments, especially when they are frequently repeated, we tend to become more impervious to others' reasons (Barberá et al. 2015; Levendusky 2013; Sunstein 2017; Unkelbach et al. 2019; Vicario et al. 2016). Hence, according to this theory, polarization occurs after arguing with likeminded people because of the exposition to a limited and skewed pool of arguments that reinforces, by repetition, our initial positions. This seems to be one of the mechanisms behind some peculiar situations of public disagreement that fuel polarization, public crossed disagreements (see section 4.5 and section 9.2).

The third major approach to the phenomenon of likeminded deliberation is social identity theory (Mackie 1986; Tajfel 1970; Tajfel & Turner 1979; Turner 1981). According to this view, our social identities, which are responsible for polarization within a group of likeminded people, are construed within the social group we belong to, and in opposition to other groups. That is to say, our identities, i.e., who we are, are fixed by opposition to other groups, especially by emphasizing the characteristics and opinions that define us and distinguish from the rest. Our wish to preserve our identity leads us to minimize intragroup differences. Hence, if I am a member of the group X, I am prone to shift my opinion in the direction of the opinion maintained by the group because of my psychological need to preserve my own identity. We will come back later to this mechanism (section 4.2.1).

The main tenet of the final approach we discuss here, i.e., corroboration theory, is that there is a close link between confidence and corroboration by others. The idea is pretty simple and intuitive: the more people agree with my opinion, the more confident I become of it. Agreement, according to this view, works as a reinforcement of prior beliefs. If I believe that p, and most people I discuss with agree with that, then my confidence in p is reinforced. This is how corroboration theory explains this mechanism: polarization occurs as a consequence of being corroborated by likeminded people. Inasmuch as one becomes more confident in one's beliefs, one becomes more extreme in one's beliefs. This link between confidence and polarization has been explicitly pointed out by some authors: "people with extreme views tend to have more confidence that they are right, and that as people gain confidence they become more extreme in their beliefs" (Sunstein et al. 2006: 75; see also Sunstein 2017). Let's make two general remarks about these theories.

First, it is important to note that these four theories point in the direction of four mechanisms that can promote political polarization not just in a likeminded deliberation situation. That is to say, our need to be socially accepted by others, or the persuasive role of arguments in favor of a preexisting opinion, especially when the pool of arguments is limited and skewed, are general mechanisms that can foster polarization without the need to discuss with likeminded people. For instance, I can be exposed to a limited and skewed pool of arguments not because I discuss with likeminded people, but due to the press I read. And the same goes for social identity and corroboration. Thus, although they are four explanations of why polarization after discussing with likeminded people occurs, they are not restricted to it. My desire for social acceptance and my need to preserve my social identity can polarize me without undergoing a process of deliberation with likeminded people. Likewise, my previous beliefs may be corroborated, and I may be exposed to a limited and biased set of arguments, without undergoing a process of deliberation with likeminded people.

Nonetheless, the general nature of these mechanisms should not detract from the fact that likeminded deliberation is a strikingly widespread phenomenon, and therefore it is strongly relevant to the matter of how we polarize. Discussing with likeminded people, on whatever issue, usually polarizes us. Some authors, such as Talisse, have recently argued that, despite the fact that political polarization and likeminded deliberation are two distinguishable phenomena, the toxic division that many contemporary democracies face is the result of likeminded deliberation in an overpolitized setting (Talisse 2019: 98; see also Aikin & Talisse 2020).

Second, It should be noted that the shift in opinions produced after likeminded deliberation typically occurs unconsciously, or at least not deliberately (Talisse 2019: 105). These processes often go unnoticed. And that is not surprising. It is common not to know exactly what we believe in a lot of issues (see section 3.4.1). In fact, we mostly discover and shape our beliefs while arguing with other people. In other words, we can hardly know if we already had the opinions we have, have acquired them during the last discussions, or if they have been at least partially modified (see chapters 5 and 6 for a deeper discussion on beliefs and other attitudes related to polarization). Then, we cannot easily identify whether we are experiencing a likeminded deliberation process.

Likeminded deliberation is not a process in which a subject undergoes conscious changes. Nothing could be further from that. Rather, it is in line with our normal belief formation processes. We continually shape our mind through social contact and discussion with other people. We constantly and jointly build our minds. Our beliefs are formed, so to speak, on a jointly built structure, on a river-bed, as Wittgenstein calls it (Wittgenstein 1969 § 97), that nonetheless keeps in continuous change. Deliberating with people from our group is just one way in which we build the structure on which many of our other beliefs rest, and that's why it goes unnoticed. In fact, discussing with people from our own group is an extremely effective evolutionary mechanism; the problem of this phenomenon, and the different mechanisms involved in it, begins when social conditions are such that they generate a harmful and dangerous division (Mercier & Sperber 2017).

In what follows we will discuss the question of what it means to say that our positions become *more extreme* as a result of a deliberation process with likeminded people. In section 2.5, following Talisse (Talisse 2019; see also Aikin & Talisse 2020), we have already introduced the distinction between belief content and degree of belief regarding the possible consequences of a situation of this phenomenon. Talisse wonders about the sense of *extreme* in which this phenomenon does render us more extreme.

We believe that the resulting extremization of deliberating with likeminded people is more naturally explained in terms of radicalism rather than in terms of extremism, even though both outcomes are conceptually possible, and the results of various experiments can be explained in both senses. We have two main reasons to embrace this. The first one is that, understood as extremism, likeminded deliberation can hardly be explained through persuasive arguments theory and corroboration theory, while all theories seem correct if it is understood as radicalism. The second reason is that many of the empirical studies on this phenomenon, if not all, use scales through which participants have to rate how much they agree or disagree with a claim, and shifts in these responses seem to indicate a change in the degree of belief rather than in the content believed.

Let's discuss the first reason. According to the theory of social comparison and the theory of social identity, polarization within a group of people that think likewise occurs because our desire to be accepted by the members of the group we belong to, and because we want to preserve our social identity. In this sense, it seems correct to say that if the group of people I interact and identify with has the belief that p, then I will adopt a belief with a slightly more extreme content for preserving my identity and my social reputation. And something similar goes for the degree of belief: if my people believe that p, then I will increase my confidence in that belief, leading me for example to manifest more extreme behaviors to preserve my identity and to be positively perceived.

However, extremism does not seem to fit with the explanation of the two other theories. Let's see why. According to persuasive arguments theory and corroboration theory, polarization occurs because in arguing with likeminded people we are exposed to a limited and skewed pool of arguments that reinforces our initial beliefs, and because the resulting agreement itself also increases our confidence in what we already believed. While it is obvious why agreement and exposure to a limited set of arguments that favors our initial position might increase our degree of belief, it is not at all obvious why these situations would necessarily lead us to believe more extreme contents. New arguments and information supporting my position, and the resulting agreement, reinforce what I already believed. For instance, if I believed that the tables of a house should be round, then I will be more confident on it as a result of being exposed to a constant agreement and a limited set of arguments that favor my position. But it is not at all obvious why this will lead me to believe, for example, that tables should be round not only in a house but in all places, or another more extreme content in this line.

Another very different thing is what happens when the belief in which we have increased our confidence is a core belief of our identity. For instance, if you believed that the measures taken by the government to stop the spread of the coronavirus were insufficient, and this is a core belief of your political identity, then it seems more convincing to say that, as a result of being exposed to a particular pool of arguments and to be corroborated, you ended up believing that the coronavirus pandemic is the government's fault. But not only this. Also that you ended up protesting with a saucepan and a golf club while yelling that this government is genocidal, and carrying out other shocking behaviors that are the result of the high level of credence in that particular belief. In this case, the simple increase in the degree of belief makes us ardent supporters of our position, as Talisse says (see section 2.5), and this can be also manifested in the adoption of beliefs whose contents are more extreme than the initial ones. But this is not a simple consequence of being exposed to a limited and biased set of arguments; it is just a *possible* consequence of giving more credibility to a core belief of our identity. Therefore, the claim that discussing with likeminded people makes us more extreme, in the sense of belief content, does not naturally fit with the mechanisms involved in the persuasive arguments theory and the corroboration theory.

Let us now discuss the second reason. If we carefully review the experiments conducted on likeminded deliberation, most of them measure the shifts in participants' responses in terms of agreement and disagreement with a claim. For instance, in a well-known study about social corroboration and opinion extremity, participants had to rate a set of photos using a scale from 0 (extremely unattractive) to 11 (extremely attractive), to assess how comfortable or uncomfortable a dentist was using a scale of -50 (extremely uncomfortable) to 50 (extremely comfortable), and to indicate how much money would they donate to a medical organization (Baron et al. 1996). Variations in responses on these scales do not seem to indicate a change in the content of belief, but rather a change in the degree of belief. By varying their response after arguing with likeminded people, participants are indicating how confident they are in believing that a certain photo is attractive, that a certain dentist was comfortable, or that a particular medical organization does a great job. And the same is true for the experiments previously discussed in this section. Thus, the very design of these experiments suggests that what these experiments really measure is radicalism rather than extremism.

In sum, deliberation within a group of likeminded people is a phenomenon that causes polarization. Although it does not necessarily assume a particular understanding of polarization and is therefore compatible with both ways of conceiving the division that characterizes political polarization, it does seem to be more akin to the understanding that has to do with the degree of belief than with the belief content. Understanding the polarization resulting from this phenomenon as extremism does not naturally fit with two well-established theories of the phenomenon, and makes it difficult to explain how most experiments actually measure this kind of shift.

## 4.2. Belief polarization: Mixed evidence disagreement

The question of how disagreement affects polarization has been widely studied from social psychology, and has received some attention from philosophy too, especially from the literature on epistemology of disagreement (see, for instance, Kelly 2008). In particular, there seems to be a specific phenomenon crucially linked to the rise of political polarization. This phenomenon consists in situations where two individuals who disagree on a particular issue, and therefore have contradictory or incompatible attitudes about it, adopt more extreme positions after being exposed to the same mixed body of evidence. For convenience, henceforth we will call this phenomenon *mixed evidence disagreement*.

Recall the example of Dereca and Izquerri: Izquerri believed that the policies adopted by the Spanish government to control the spread of the COVID19 were reasonable, and Dereca believed that the policies adopted should have been other, and therefore that they were not reasonable. In this situation, Dereca and Izquerri disagreed on whether the policies taken by the government were appropriate. Suppose next that both were exposed to the same body of evidence, which consists of two sets of data, i.e., E1 and E2. Imagine that E1 refers to a data set that shows that the policies taken by Spain were reasonable because they have been more effective in saving lives than the measures taken by many other countries. E2, on the other hand, refers to a set of data that shows that the policies adopted by

Spain were not reasonable because they have had a greater negative impact on the economy than the measures adopted by many other countries. Imagine that Izquerri considers E1 to be much more convincing and probative than E2, while Dereca thinks exactly the opposite. In this case, Izquerri reinforces her initial belief, and the same happens with Dereca. As a consequence, both hold more extreme positions in the direction of their initial ones, and therefore disagree more harshly than they did at the beginning. This example grasps the basic idea of the phenomenon we want to discuss here: two people who disagree tend to polarize after being exposed to the same mixed body of evidence.

As Carter and Broncano-Berrocal note, at least three differences can be drawn between mixed evidence disagreement, which they call 'belief polarization', and likeminded deliberation, which they call 'group polarization' (Broncano-Berrocal & Carter 2021). First, since mixed evidence disagreement has to do with how a given body of evidence affects the positions of two individuals who hold opposite positions, this phenomenon is not about collectives, i.e., it does not have to do with processes that occur within groups of people who discuss a particular issue, but with processes *individually* experienced when subjects are exposed to certain evidence. Second, unlike likeminded deliberation, mixed evidence disagreement arises without deliberation. That is to say, the *absence of deliberation* is one of its defining features. Third, mixed evidence disagreement seems to arise from a *prior disagreement* between two individuals, while likeminded deliberation departs from a prior agreement, i.e., after deliberating with likeminded people. These are three defining differences between both phenomena.[1]

The label 'belief polarization' as a label to refer to mixed evidence disagreement was coined by Lord, Ross, and Lepper in the late 1970s (Lord et al. 1979). These scholars hypothesized, against the idea that increasing the amount of evi-

[1] Regarding the third difference, we would like to note that, in the case previously considered as an attempt to introduce the phenomenon of being exposed to a mixed body of evidence, and surely in most cases of it, the disagreement involved is *deep disagreement*. That is to say, a kind of disagreement in which each part has a different standard from which they judge (Lynch 2010; Smith & Lynch 2020). For example, each part considers a different principle or method as the most relevant to resolve the disagreement.

dence on which a decision is based tends to alleviate existing discrepancies, that a given disagreement tends to deepen when both sides are exposed to the same body of mixed evidence. To test the hypothesis, they conducted an experiment with 48 undergraduate students carefully recruited. 24 of these students were advocates of capital punishment, believed that such punishment has a deterrent effect, and maintained that relevant research supports their position. The remaining 24 participants were opponents of capital punishment, questioned its deterrent effect, and believed that the relevant research supports their position too. The procedure was the following. Participants were shown the results of two invented but realistic studies. According to one of them, the murder rate was lower in 11 of 14 states the year after the death penalty was adopted. According to the other study, after comparing murder rates in 10 pairs of neighboring states with different capital punishment laws, the results showed that murder rates were higher in the state with the death penalty in 8 of the 10 pairs. After being exposed to both studies, controlling a possible order effect, participants were asked to respond whether they were more opposed or more in favor to capital punishment, and if they were less convinced or more convinced that capital punishment has a deterrent effect, using a scale from -8 to 8. In addition, they were also asked to judge how well each study had been conducted, and how convincing each study appears on the deterrent efficacy of the death penalty, using the same scale. The results confirmed the hypothesis of the experimenters. Participants who initially favored that death penalty has a deterrent effect evaluated more positively the study that agreed with them, and reinforced their position, i.e., their attitudes became more extreme in the direction of the initial position.

According to the authors, the resultant polarization responds to the biased way in which we are prone to evaluate a body of mixed evidence. As Lord, Ross and Lepper put it, “This ‘polarization hypothesis’ can be derived from the simple assumption that data relevant to a belief are not processed impartially. Instead, judgments about the validity, reliability, relevance, and sometimes even the meaning of proffered evidence are biased by the apparent consistency of that evidence with the perceiver’s theories and expectations” (Lord et al. 1979: 2099). This polarizing effect has been widely replicated across a range of topics (Chen et al.

1992; Hastorf & Cantril 1954; Houston & Fazio 1989; Kunda 1987; Koehler 1993; Liberman & Chaiken 1992; Munro & Ditto 1997).[2]

Recent findings seem to support this idea that we don't only get polarized inside homogeneous groups, but also when we get exposed to the evidence and reasons supporting the view of "the others".[3] In particular, they seem to support the idea that we process information in a biased way. For instance, it appears that forcefully follow a Twitter bot that retweets information from the other political side makes us more polarized, at least when we are openly affiliated to a political identity (Bail et al. 2018). In addition, it seems that correction of misleading information creates a "backfire effect", which increases our misperceptions and produce more polarization instead of less (Nyhan & Reifler 2010). Being exposed to information supporting others view, even when it is accurate factual information that corrects our misperceptions, seems to foster polarization. The idea would be similar to the way a detective interrogates a person she considers a suspect: every statement made by the person under interrogation may be reinterpreted in a way compatible with that the suspect is guilty, and thus reinforcing the detective's previous belief.

However, the story of the causes of this phenomenon is not necessary one characterized by our biased way of processing information supporting the other

[2] It is important to note that some studies have partially challenged the phenomenon of mixed evidence disagreement (Miller et al. 1993; Munro & Ditto 1997). According to these studies, when the distinction between self-reported and directly assessed attitude is taken into consideration, results vary. If participants are asked to self-report the change of their initial attitude, i.e., whether their attitudes had become more favorable or not toward p, then mixed evidence disagreement was found. But if participants' attitudes were analyzed directly by comparing pre- and post-attitude responses, then the results do not show mixed evidence disagreement. We note this here because the distinction between self-reporting and directly assessing people's attitudes, with some adjustments, is crucial to this dissertation, as we will see in the following chapters. Otherwise, the criticism posed by these studies, although interesting, does not challenge the pervasiveness of mixed evidence disagreement.

[3] It is important to note, however, that recent research suggests that the outcome of getting more polarized, or on the contrary reaching agreement, as a result of a discussion crucially depends on the particularities of the members involved in it. More specifically, if most members involved in a disagreement have low confidence in their views, then the group is more likely to succeeding at reaching consensus (Navajas et al. 2019). This is an important finding.

side's view. Some researchers have suggested that our reaction to ambiguous information is not to reinterpret it to make it more coherent with what we already believe, but to scrutinize it closely, and the more the scrutiny, the more likely to find a flaw in that argument, which reinforces our initial belief (Gilovich 1991; Kelly 2008). In fact, according to some recent findings, the "backfire effect" is not widely replicated (Wood & Porter 2019): correction of misleading information does make us more accurate in our factual beliefs. But when we correct our misperception, the correction itself does not affect to our political preferences (Porter et al. 2019). So, according to this second trend, the effect of these situations in which two people in disagreement are exposed to a mixed body of evidence is not caused by our dogmatic way of processing information, but by our close scrutiny of those arguments against our political preferences, which make us more factually accurate but does not change our previous political beliefs. The reason might be that our close scrutiny of new information increases the size and density of the set of arguments supporting our political identity's core beliefs. In this sense, the underlying mechanism seems as the same mechanism discussed in the previous section, the mechanism pointed out by the persuasive arguments theory.

As in the case of the previous one, different approaches to this polarizing phenomenon can be distinguished. In what follows we will discuss a little bit more these two allegedly competing explanations, which have been already discussed by Kelly (Kelly 2008), and argue that each of them seems quite plausible either taken in isolation, and considered in relation to the other one.

### 4.2.1. Digging into dogmatism and scrutiny models

Kelly discusses two alternative models of how individuals respond to evidence that does not support their initial beliefs (Kelly 2008). The first one, which he calls *Kripkean dogmatism*, is as follows. If I believe that p, and I'm exposed to a body of mixed evidence that both supports and does not support p, then it is rational to me to dismiss the portion of the evidence conflicting with my belief just because there is a portion of the evidence that in fact supports my belief that p. As a result, I become increasingly confident about the truth of p, and I treat my belief as a license to discount exactly that portion of the mixed body of evidence that does not support my previous belief. And you, that believe that not-p, reason exactly the same but in the opposite way.

Kelly dismisses this approach as a plausible explanation. According to Kelly, this process makes us unreasonable: if I am justified to think that any counterevidence will be misleading, I am justified in ignoring such evidence when I actually encounter it, and that is an unreasonable and dogmatic attitude. Besides, Kelly argues that Kripkean dogmatism gives too much importance to the temporal order in which a person encounters some pieces of evidence in the belief formation process, which contradicts the Commutativity of Evidence Principle: what it is reasonable for one to believe depends on one's total evidence.

Kelly offers an alternative to Kripkean dogmatism model, which we call Kellyan scrutiny. According to *Kellyan scrutiny*, we don't dismiss counterevidence in an instance of mixed evidence disagreement, but actually pay more attention to it. If I believe that p, I also believe, or at least I am willing to believe, that there are no good arguments for not-p. When an argument is presented as a convincing argument for not-p, I view this argument with a greater measure of suspicion and subject it to closer scrutiny. And the more the scrutiny, the more likely to find a flaw in that argument. As a consequence, I search another hypothesis p* as a correct explanation for the argument, and then increases the level of confidence I give to my initial belief that p. Given that it seems to be rational to us to scrutiny what does not fit with what we believe, this second explanation seems more plausible than the Kripkean dogmatism, argues Kelly. In fact, as Kelly noted, it enables us to explain the participant's reaction in the Lord, Ross & Lepper's study (see Kelly 2008: 617-619; see also Gilovich 1991).

We agree with Kelly that Kellyan scrutiny model points to a psychological mechanism that occurs in mixed evidence disagreement, at least in some contexts where there is an explicit or implicit requirement not to ignore the opposing positions and arguments. Moreover, as we have seen above, recent findings support this hypothesis (Porter et al. 2019). However, against Kelly, we do not believe that the Kripkean dogmatism model is necessarily incorrect, nor irrational. It is not difficult to imagine a situation where we would say that it is rational not to pay attention to the 'evidence' that contradicts our initial belief. For instance, studies whose results try to confirm the hypothesis that there is no global warming, that the Earth is flat, or that a certain population is less intelligent by nature, can be ignored for many different reasons, at least in some contexts, without being irrational, even if one does not know in detail the arguments against the outcome of these studies. This is so partly because, in some contexts of our daily life, we would not say that it is irrational to trust the conclusions advocated by certain groups, such as the scientific community or others, to the point of not paying too much attention to arguments that go against them. It has many evolutionary advantages and fits well with some of the mechanisms discussed in the previous section.

Besides, to say that something or someone is irrational is usually to make an evaluation, and evaluations are always relative to a standard, on which their truth or falsity depends (Field 2009; Frápolli & Villanueva 2018; Gibbard 1990, 2012). Depending on the standard, and the importance attached to each of its parameters, what a person does can be deemed rational or irrational, and there is no matter of the fact that necessarily settles the dispute (Frápolli & Villanueva 2018), at least when it is evaluatively used.

Regarding the understanding of polarization that mixed evidence disagreement presupposes, it appears that, as happened with likeminded deliberation, this phenomenon does not necessarily assume a particular understanding of polarization. That is to say, after being exposed to a body of mixed evidence, one can polarize in the sense of adopting beliefs with more extreme contents, or by increasing the degree of confidence in the previous beliefs. However, as we have seen, mixed evidence disagreement has to do with *reinforcing* the initial position after being exposed to mixed evidence. Participants often have to respond whether they are more or less *convinced* after reading some studies, when mixed evidence disagreement is measured through self-reports, or by indicating if they *strongly agree or disagree* with some claims before and after being exposed to two studies with contradictory results, when their attitudes are 'directly' measured. As we have already discussed in the last section, this way of measuring attitudes seems to indicate shifts in degree of belief rather than in belief content. Therefore, mixed evidence disagreement appears to be explained in a more natural way by understanding the resultant polarization in terms of radicalism.

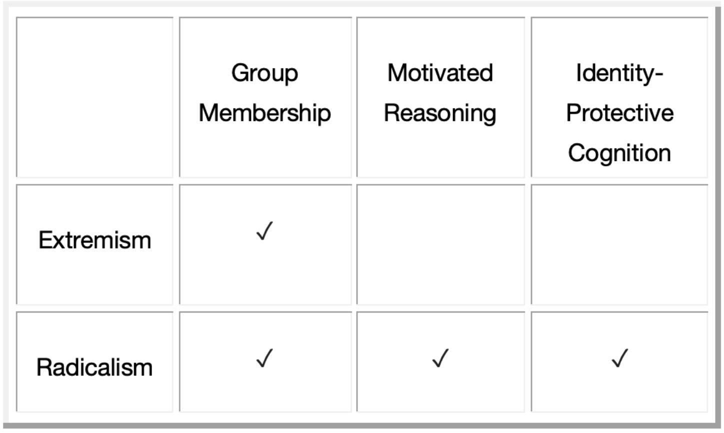
## 4.3. Psychological phenomena fueling polarization

The aim of this section is to review three general mechanisms, that can be deemed as psychological ones, that bring about political polarization. It is important to keep in mind that so far, in this chapter, we have discussed two well-known and replicated phenomena closely related to the rise of political polarization (like-minded deliberation and mixed evidence disagreement), have stressed four general mechanisms also involved in the rise of polarization (our need to be accepted by members of our group, our need to preserve our identity, the role played by the arguments we are exposed to in certain contexts, and the role played by corroboration), and have highlighted two attitudes that may be related to the increase of polarization under certain conditions (dogmatism and scrutiny). Some of the mechanisms that will be reviewed and discussed in what follows include some of the previously discussed ones. In particular, in this section we discuss the mechanisms of *group membership*, *motivated reasoning*, and *identity-protective cognition.*

*Group membership* affects the way we see in-group and out-group people. In-group's abilities, arguments, and reasons are seen under a more positive light than those of out-group people, simply because they came from members of our own group. The in-group favoritism is easily construed; the simple split into two groups is sufficient for it, no matter the reason for the division (Billig & Tajfel 1973). This is known as the *minimal group paradigm* (see Mason 2018). This mechanism is not necessarily negative. Group formation has evolutionary and psychological advantages (Tajfel et al. 1971). As Brewer argued, humans have two basic social needs: the need to be part of a group and the need to build its identity in opposition to others. By creating groups that are distinguishable from others we satisfy both needs (Brewer 1991). The problem with groups creation, however, arises when it leads to a neat separation between them. Our various identities often intersect with each other. For instance, two racial identities can cluster into the same political identity, split in relation to sport or music identities, and cluster again in geographical identity. However, social groups become dangerous when they involve cutting social ties with people outside the group. That is to say, when social identities are sharply aligned within a group (e.g., racial, religious, geographical, etc.), it becomes dangerously separated from other groups, due to the mechanism of favoring in-group people. Mason calls *mega-identities* to the group resulting from the fusion of different social identities into a single political group (Mason 2018: 43). This mechanism is closely related to comparison theory and identity theory about likeminded deliberation. At first glance, it seems that group membership might promote both extremism and radicalism: our tendency to favoring our in-group members may lead us both to adopt particular belief contents and certain degree of belief.

*Motivated reasoning* is an information processing mechanism. The basic idea is that we humans are often prone to process information in a biased way. Motivation affects reasoning. In particular, it seems that we tend to produce justification in favor of a preferred conclusion (Hetherington & Rudolph 2015: 77-79; see also Kunda 1990; Taber & Lodge 2006). Thus, a motivated belief or desired outcome functions as a filter that affects our evaluation of the evidence and arguments we encounter. Evidence and arguments supporting our desired conclusions are seen as more conclusive than those that contradict them. A similar but different notion is the notion of confirmation bias, sometimes conflated in the literature. *Confirmation bias* is the tendency to seek information that supports a prior position, and interpret evidence and information in ways that favor preexisting beliefs (see Nickerson 1999 for a review of the phenomenon). Sometimes, confirmation bias is deemed as a product of motivated reasoning (see Sunstein 2017): our motivation to favor a certain conclusion leads us to seek and interpret information in a biased way. However, it could happen that we seek and interpret information in a biased way without the desire to favor a certain conclusion. In this sense, both mechanisms are separable. These mechanisms are related to the phenomenon of mixed evidence disagreement, more specifically to the dogmatism model (section 4.2.1). But it also appears to be involved in likeminded deliberation to the extent that seems to point to our tendency to discuss with people that think like us. According to Sperber and others (Sperber et al. 2010), we humans possess some cognitive mechanisms for epistemic vigilance that buttress our mutual trust when sharing information through communication. However, in some social settings, the effect of epistemic vigilance dismisses: we tend to accept the information that comes from members of our cultural group much more frequently than when it comes from out-group people (Sperber et al. 2010: 380-381). Evidence search and interpretation in favor of a prior position seems to foster radicalism rather than extremism, so it seems that motivated reasoning is prone to bring about polarization in terms of degree of belief.

*Identity-protective cognition* is a form of motivated reasoning, which differs from it in that the goal of the reasoning is one of a specific type: to protect one's own status, by promoting information that favors the cultural commitments of the group in which the defended position is embedded. The idea is that if an individual defends that there is no global warming, that individual will pay more attention to the information whose defense expresses her commitments to the position that there is no global warming (Kahan et al. 2011; Kahan 2017). For example, that individual will not consider as an expert a particular scientist who defends global warming; and will agree that the same scientist, with the same credentials, is an expert if she defends that there is no global warming (see Kahan et al. 2011). Identity-protective cognition can be seen as a mix of group membership and motivated reasoning. Given that identity-protective cognition is a kind of motivated reasoning, it seems that this mechanism promotes radicalism rather than extremism. The following table shows the compatibility of the mechanisms with extremism and radicalism.



| | Group Membership | Motivated Reasoning | Identity-Protective Cognition |
|---|---|---|---|
| Extremism | ✓ | | |
| Radicalism | ✓ | ✓ | ✓ |

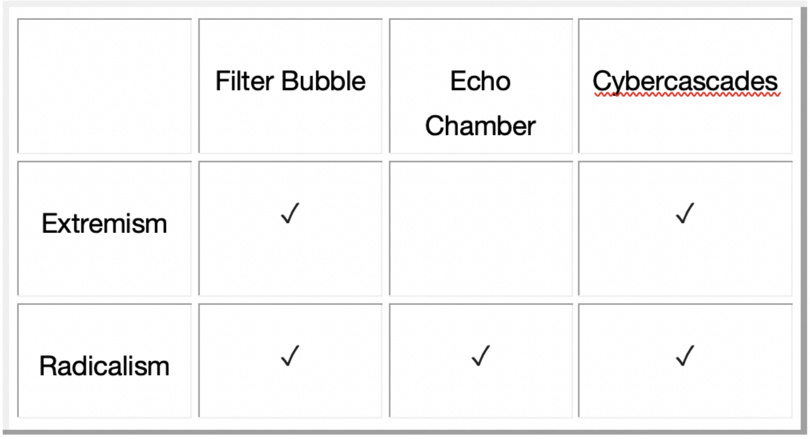
## 4.4. Social phenomena fueling polarization

The aim of this section is to review three general mechanisms promoting political polarization that can be deemed as social ones. In particular, in this section we discuss the mechanisms of *filter bubbles*, *echo chambers*, and *cybercascades*.

*Filter bubbles* are nondeliberately personalized universes of information. In a sense, it has many resemblances to epistemic bubbles and echo chambers, as we will see in what follows. But it refers to a slightly different phenomenon. Given the prediction engines of the Internet, based on our individual behavior online, each of us is exposed to information in a very individualized way and, as a result, is in a particular 'universe' (Pariser 2011). Your location, the things you like, search for, listen and watch through several platforms are constantly refining your 'profile' to filter for you what you like (see Tufekci 2017). As a result, each of us is in a filter bubble, a finely tuned personalization. Pariser highlights three features of filter bubbles. First, each person is the only person in her bubble. Second, the specific reasons, i.e., the choices made by the algorithms, from which a particular content is offered, are not transparent to us. Third, our personalized universe is not entirely the result of our decision (Pariser 2011: 11-12). It seems that filter bubbles can promote both radicalism and extremism by offering information that reinforces previous belief and by offering information in favor of more extreme contents in line with a prior belief. However, given that the finely tuned personalization consists in offering to you the things you already like and agree with, it is reasonable to think that, as an outcome, people get more radicalized because their confidence in their previous beliefs increases.

*Echo chambers*, sometimes called as information cocoons (see Sunstein 2017) are spaces that have the power to magnify the messages shared within it and protect them from refutation, generating a positive feedback loop for those exposed to the messages shared in that space (Jamieson & Cappella 2010: 76; see also Bordonaba 2020: 305; Sunstein 2017). In an echo chamber, the possibility of being exposed to contrary ideas is extremely limited. As a result, it exaggerates their member's confidence in their beliefs. For example, a television channel, or a Twitter account, can function as an echo chamber inasmuch as it amplifies only certain opinions. In echo chambers, an individual is just exposed to likeminded people's opinions. According to Nguyen, one of the defining features of echo chambers is not just the exclusion of certain sources of information, but the actively discredit of dissenting voices (Nguyen 2020). He emphasizes this feature to differentiate echo chambers from epistemic bubbles, a phenomenon that is usually lumped in with echo chamber. *Epistemic bubbles*, like echo chambers, are social structures that subvert the flow of information by excluding certain ideas. However, unlike echo chambers, in epistemic bubbles the dissenting voices are not actively undermined, but excluded by omission (Nguyen 2020). The omission can be purposeful if the exclusion is selectively avoided, for instance by blocking through Twitter those with contrary views. But it can be also inadvertent, for example because our Facebook friends only share information that fits with our view and we are not aware of that. Both phenomena, echo chambers and epistemic bubbles, pave the way for reinforcing the preexisting views because it affects to the density and size of the pool of arguments which people are exposed to, and, in this sense, promote polarization (see Barberá et al. 2015; Sunstein 2017; Vicario et al. 2016). Hence, it seems that these phenomena promote radicalism rather than extremism (see Almagro & Villanueva 2021).

*Cybercascades*, also called social cascades, information cascades, or just cascades, are flows of information exchange in which certain information, including false information, becomes widespread through social media simply because many people share and endorse it (Sunstein 2017). That is to say, a crowd supporting a piece of information is usually seen as an extra reason to believe in the information conveyed. As an example of cybercascades, take the case that appears in (Lynch 2016: 78): hundreds of thousands of people retweeted the picture of a man holding a wounded woman the day of the bombing of the Boston Marathon in 2013, together with a message telling a (false, as later became known) story: the man had planned to propose to the woman at the end of the marathon. Sunstein distinguishes two kinds of cascades. Informational and reputational cascades. The former occurs when people share an opinion simply because the number of people who consider it to be true provides extra confidence. The latter occurs when people go along with the crowd because of the pressure, even though they don't actually endorse the content shared (Sunstein 2017). Endorsing a piece of information simply because is supported by a crowd seems to be able to promote both extremism and radicalism. The following table shows the compatibility of the mechanisms with extremism and radicalism.



| | Filter Bubble | Echo Chamber | Cybercascades |
|---|---|---|---|
| Extremism | ✓ | | ✓ |
| Radicalism | ✓ | ✓ | ✓ |

## 4.5. Linguistic phenomena fueling polarization

The aim of this section is to review three general mechanisms promoting political polarization that can be deemed as linguistic ones. In particular, in this section we discuss the mechanisms of *abstract vs. concrete uses of language*, *dogwhistles*, and *crossed disagreements*.

*Abstract and concrete uses of language*. According to several empirical studies (see Napier & Luguri 2016), the abstract and concrete use of language influences the positive and negative feelings of liberals and conservatives toward certain groups, and toward certain policies. On the one hand, it seems that when conservatives and liberals think in concrete terms (e.g., ringing a doorbell as pushing a button) vs. abstract terms (e.g., ringing a doorbell for seeing if someone is in home), and then are rate their positive or negative feelings toward non-normative groups (homosexuals, atheists, Muslims), polarization between conservatives and liberals increases regarding their prejudices toward non-normative groups. While in the abstract mindset, polarization is reduced (Napier & Luguri 2016: 146-152). This outcome is replicated not just when the abstract mindset is spontaneous, but also when it is induced by the experimenters. On the other hand, it seems that if liberals and conservatives are induced to think in abstract terms (vs. concrete terms) and evaluate some contemporary political issues, polarization between them increases when partisan identity is salient, and reduces when national identity is salient (Napier & Luguri 2016: 153-155).

When political identity is salient and we are induced to think in abstract (vs. concrete) terms, the distance regarding the positive or negative feelings generated by certain concrete policies expands between conservatives and liberals. In particular, conservatives tend to evaluate these policies more negatively when they think about it in abstract terms. Conversely, when national identity is salient, thinking in abstract (vs. concrete) terms reduces the distance between the positive and negative feelings of conservatives and liberals. Conservatives tend to indicate more positive feelings. These results are in line with the findings of other studies, according to which when national identity is salient, the level at which Democrats dislike Republicans, and vice versa, falls (Levendusky 2018). However, when national identity is salient, the level of dislike toward people forced to move out of their origin countries increases (Wojcieszak & Garrett 2018). Since these empirical studies measure positive and negative feelings using the tool of feeling thermometer, ranging from 0 to 100, and using a scale from 0 (strong negative feeling) to 9 (strong positive feeling) in the case of policies, it seems that they indicated radicalism rather than extremism, at least in certain situations (see section chapters 7 and 8).

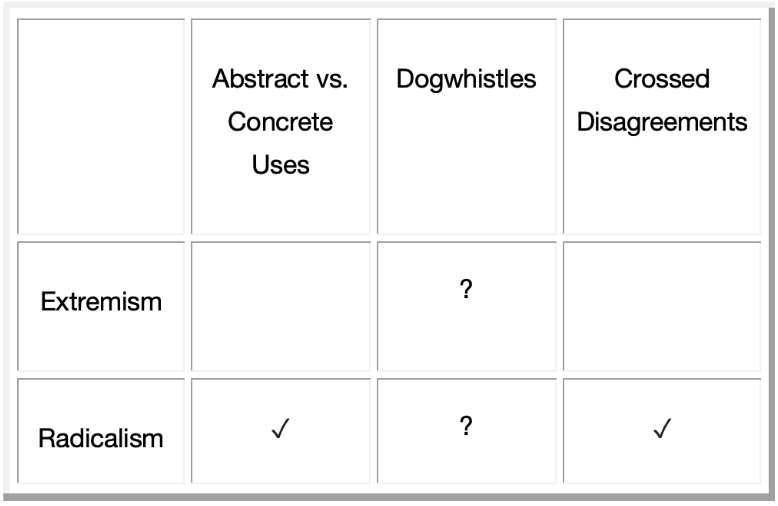
*Dogwhistles* are a particular strategy of propaganda (Stanley 2015), usually employed in politics, to obtain certain benefits in a covert way. In particular, they are speeches that employ codified language and particular expressions to convey additional information to a subgroup of the general public that passes unnoticed by the rest of the audience (Khoo 2017; Mendelberg 2001; Saul 2018; Stanley 2015; Torices 2019, Torices forthcoming). Dogwhistles can be intentional or unintentional, and overt or covert. The intentionality feature has to do with whether the speaker *deliberately* performed the dogwhistle. Since encoded information often depends exclusively on the use of certain expressions, one can unintentionally make a dogwhistle. The distinction between overt and covert, on the other hand, has to do with whether the target group is aware about the hidden message and can decode it. When the target group consciously recognize the hidden message, the dogwhistle is overt. When the information conveyed is not retrieved and not consciously recognized by the target group, but still has a certain effect and promotes certain attitudes in members of that group, then the dogwhistle is covert (Saul 2018: 365). Frequently, the coded information is about speaker's sympathy for a certain policy or ideology when overt, and promotes racist and exclusionary attitudes toward people from certain social identities when covert. In this sense, it seems that, when dogwhistles promote certain attitudes, they promote radicalism rather than extremism, because attitudes seem to be the expression of having a particular level of confidence in a particular opinion rather than having a particular opinion (insofar as one could believe certain content and not express certain attitudes, like a non-polarized racist, as we have shown in chapter 1). The effectiveness of covert dogwhistles depends on the existence of previous attitudes, for example racist ones. So it makes sense to think that, after being exposed to a covert dogwshistle, the target group will more strongly hold an opinion they already had, and in this sense dogwhistles would promote radicalism rather than extremism. But in the absence of empirical studies, we leave this question open here.

*Crossed disagreements* are situations in which each party involved in a disagreement clearly exhibits signs of conceiving the disagreement in significantly different terms (see Almagro et al. 2021; Bordonaba & Villanueva 2018; Osorio & Villanueva 2019). When two sides are discussing a given topic, the disagreement might be about a matter of fact or, on the contrary, about a normative issue. When a dispute is about a normative issue, the discussion may be about the standards that each party relies on to uphold its position –and therefore the dispute can become about the standard to be adopted. But the discussion may also be about the issue itself, and not about the standard to adopt, e.g., a discussion about whether a particular action is good or bad. The former can be called normative disagreement, while the latter evaluative disagreement. So, a disagreement can be factual, normative, or evaluative.[4] The crucial difference between factual and non-factual disagreements is that, in the former, both parts share their standards, and if they discover that they don't share the same standard, then the discussion becomes meaningless, or it becomes about the standard that should be adopted. That is to say, in a factual disagreement, the disagreement cannot persist after discovering that each party has a different standard without turning out in a discussion about the standard. However, this could happen in a non-factual disagreement: after making explicit that both parties have different standards, the disagreement can persist without turning into a discussion about the standard that should be adopted.

Besides, there can be a dispute in which, even if the speakers are not talking at cross purposes, they are involved in a crossed disagreement inasmuch as they are displaying clear signs of conceiving the disagreement as if it were of a different nature. That is to say, maybe one side, given the arguments adduced, is clearly conceiving the disagreement as a factual one. The other part, on the

[4] Disagreements can be of many other kinds. In the literature about disagreement, there had been distinguished different kinds: faultless disagreement (Kölbel 2004; Lasersohn 2005; MacFarlane 2014), metalinguistic disagreement (Sundell 2016), peer disagreement (De Cruz & De Smedt 2013), deep disagreement (Fogelin 1985; Lynch 2010; Smith & Lynch 2020), or non-straightforwardly factual disagreement (Field 2009). However, following Bordonaba, Osorio and Villanueva (Bordonaba & Villanueva 2018; Osorio & Villanueva 2019), we think they can be simplified to three types of disagreement to certain purposes like the ones of this dissertation.

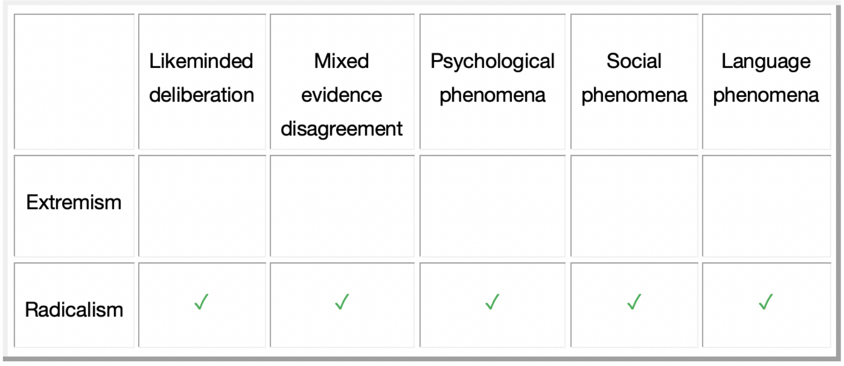
contrary, may be conceiving it as a normative disagreement, and therefore argues about the standard it should be adopted. When a situation as such occurs, the dispute counts as an instance of crossed disagreement. One of the features of a situation of crossed disagreement is that it is quite difficult to bring positions closer. Supporters of both parties involved in a crossed disagreement presumably will end up with their prior position reinforced because, given the nature of the arguments provided by each part, they will be hardly compelled by them (see Osorio & Villanueva 2019: 126). In these situations, each party is mostly exposed to arguments that reinforce their initial position. As a consequence, the repetition of the arguments that support their initial opinions makes them more convinced that the other side was wrong from the beginning. The arguments offered by one side do not really affect to the opposite side, and then each party will be just exposed to repeated arguments that support their initial belief. Thus, this sort of situation increases the size and density of the pool of arguments to which each one is exposed, which at the same time increases polarization because it increases our level of confidence in our previous beliefs (see sections 4.1 and 4.2.1). In fact, a recent study, based on the Minutes of the Sessions of the Spanish Parliament (see Bordonaba & Villanueva 2018), has shown that there is a strong correlation between the increase of crossed disagreements and polarization. The following table shows the compatibility of the mechanisms with extremism and radicalism.



| | Abstract vs. Concrete Uses | Dogwhistles | Crossed Disagreements |
|---|---|---|---|
| Extremism | | ? | |
| Radicalism | ✓ | ? | ✓ |

## 4.6. Conclusion

In this chapter we have reviewed some of the relevant evidence and some of the most established mechanisms involved in the increase of political polarization. Furthermore, we have paid attention to whether the different mechanisms and the best available evidence are compatible with the understanding of polarization that has to do with belief contents or with the conception that has to do with degree of belief, namely extremism and radicalism. As a result, the following table shows which of both is in a better position to account for the evidence revisited, in terms of whether it seems more compatible with one understanding or the other.



| | Likeminded deliberation | Mixed evidence disagreement | Psychological phenomena | Social phenomena | Language phenomena |
|---|---|---|---|---|---|
| Extremism | | | | | |
| Radicalism | ✓ | ✓ | ✓ | ✓ | ✓ |

Deliberating with people of our own group, who think in a very similar way to how we do, usually polarizes us. In particular, it leads us to increase our confidence in what we already believed. But exposing oneself to the reasons given by the opposing party under certain conditions, especially when they are seen as "the others", does not help to decrease the level of polarization as well. To put it in another way, exposing ourselves to the reasons from others, under certain circumstances such as in situations of crossed disagreement or in situations where we have the capacity to filter the other information we consume, would increase our confidence in what we already believed.

As Kevin Dorst argues, there is a striking and standard story about the rise of polarization that explains it by claiming that it is the result of our irrational and biased way of processing information. Besides the problems this story encounters that we have already pointed out in section 3.2.2, it could be a wrong story for other reasons. As Dorst notes, If we pay attention not only to the psychological evidence under the paradigm of human irrationality, but to the whole best available evidence about the rise of polarization, then the rise of polarization can be understood as "the result of reasonable people doing the best they can with the information they have" (Dorst 2020). In particular, it can be argued that the major problem of polarization lies in our informational system, together with the situations where the arguments and the evidence we get are ambiguous and then it is rational to be unsure how to react to it. Our capacity to filter the information we consume, together with phenomena such as crossed disagreements, which blur public debates and facilitates that people being exposed to them end up thinking that the evidence and arguments supporting certain positions are ambiguous, rise situations where the rational attitude is to intensify the confidence in one's previous beliefs, as occurs with the phenomenon of mixed evidence disagreement. As Ophelia Deroy has recently argued (Deroy 2019; see also Battich et al. 2020), although sharing the same experiences with people who are a bit too much like us can give rise to dangerous consequences, at the same time it is crucial for being part of a shared reality. Joint action and joint attention have enormous advantages, and part of what makes them possible is sharing experiences with people who have similar goals. And as Carmona and Villanueva point out, judging together allows us to build upon our differences, which can be used to explain why intercultural communication is possible (Carmona & Villanueva ms). In that sense, sharing certain experiences and judging together is not necessarily a bad thing, nor the result of an irrational process. It might become a problem when the environment pushes us to relate much more frequently only with certain people, and when it generates situations where information exchanges are flawed. Given the information overload and our limited capacity to pay attention to it (Lorenz-Spreen et al. 2020), it might be rational not to pay attention to the arguments coming from the other side. We will say something else on this line in chapter 8.

In the next chapter, we will start to discuss the issue of mental self-ascriptions. In particular, we will discuss whether the view according to which our attitude self-ascriptions accomplish a descriptive function is compatible with the operational notion of polarization that we are seeking along this dissertation.

# Chapter 5

# Ascribing a Mental State: Descriptivist approaches

> When we do philosophy we are like savages, primitive people, who hear the expressions of civilized humans, put a false interpretation on them, and then draw the queerest conclusions from it. (Wittgenstein PI § 194)

There seems to be a paradox concerning mental state self-ascriptions. On the one hand, it seems that speakers exhibit a presumptive authority regarding the truth of their mental self-ascriptions: if they sincerely and spontaneously claim that they believe or feel that p, then it is highly intuitive to consider their mental self-ascription as true. As Finkelstein puts it, “If you want to know what I think, feel, imagine, or intend, I am a good person –indeed, usually the best person– to ask” (Finkelstein 2003: 9). On the other hand, however, it seems that subjects frequently fail in identifying their own mental states, and even when they are self-reported in a sincere and spontaneous way, many times they are false. So, it is also highly intuitive to consider that there is a gap between what a speaker says she believes or feels and what she really believes or feels. Let’s call this situation the “mental self-ascription paradox”. Here is an example that brings out the second intuition.

> **CASE 1**: Oscar, a Spanish man, is invited to have dinner at the house of some old friends from the university. Over dinner, Oscar makes several unfortunate comments about gay people and people from disadvantaged backgrounds, and the hosts, two lesbian women from low-income neighborhoods, request him to please never make a similar comment nor use certain expressions to talk about people from certain socially oppressed groups like that, because it is highly offensive. Outraged, Oscar complains that censorship is intolerable, and claims very convinced that all people have the right to say everything they want whenever they want, anywhere. On several occasions, he sincerely self-ascribed a similar doxastic state by uttering the sentence "I believe there should be unrestricted freedom of expression". "I know what I believe, people like you are destroying democracy with this censorship. It is essential to be able to say what you want whenever you want", and so on. However, at the same time Oscar is known for being in favor of the political party that approved in 2015 in Spain the *Ley Orgánica de protección de la seguridad ciudadana*, better known as "Ley Mordaza". This law has limited the freedom of expression mainly to certain social groups with low social power. Moreover, Oscar frequently insults people who make jokes on Twitter about the Christian religion, which he professes, and threatens to report them for a hate crime. Besides, Oscar participated in a study about the offensiveness of language, and in a significant number of vignettes he responded by choosing the option 'the speaker should not say what she says'.

Does really Oscar believe that there should be unrestricted freedom of expression? We think the intuitive answer is no, despite the fact that he sincerely claimed that he does.[1] It seems intuitive to say that Oscar does not exhibit authority, even though we don't think he is lying or just trying to justify his behavior.

[1] To strengthen the claim that it is intuitively false that Oscar believes what he says he believes,

In fact, even if Oscar sincerely defends such a position frequently, it still seems intuitive to think that his sincerity does not guarantee that Oscar has the belief he says he has. We don't think that our intuition about the falsity of Oscar's self-ascription is the result of a suspicion about his sincerity or his capacity to know his mind, no matter the reason; we think that, given the whole context of the case, the presumption of authority is not triggered because his sincerity is not the most salient contextual feature in determining the truth-value of his mental self-ascription. Although fictitious, Oscar's case is not an isolated one. In a 2015 Pew Research Center survey (Wike & Simmons 2015), 71% of US Americans agreed with the statement "People can say what they want", putting the United States at the top of the list of countries that embrace the unrestricted right to freedom of expression, and 76% of Spaniards accepted that it was very important that people can say what they want without censorship in their country.

However, in many other contexts, the sincerity of the speaker does indeed guarantee the truth of her mental self-ascription, even in some more limit cases. Consider the following limit case:

> **CASE 2**: Kautar, a young British woman of Arab descent, is invited to dinner at the home of a white friend from university. The host, Kautar's friend's father, is polite and welcoming to Kautar. He is generous with the food and wine, and asks Kautar a series of questions about herself. Everyone laughs and talks amiably. As Kautar leaves, however, she is unable to shake off the conviction that her friend's father is racist against Arabs. She meets a friend to tell her about it and on several occasions she sincerely says "I believe my friend's father is a racist". Her friend presses her a bit to find out what happened, but Kautar is unable to give a reason why she thinks the host is racist. However, she explicitly states she has no doubts that she believes so. "I can feel it, and I know what I believe", she says. During

we have carried out a survey and passed it to 20 participants. Only 3 of them have the opposite intuition, namely that Oscar does believe what he says he believes.

> the conversation with her friend, Kautar says several times that she had a good time with them and wants to have dinner again another day. But Kautar is known for her strong and visceral opposition to racism. She frequently blocks people who make racist comments on Twitter, routinely stops associating with anyone she thinks is racist, and generally doesn't enjoy being around someone she suspects is racist. And yet, she insists in her impression that her friend's father is racist and wouldn't be convinced otherwise.[2]

Does really Kautar believe that her friend's father is racist? We think the intuitive answer in this case is yes, despite the fact that she wants to see her friend's father again and yet she usually hates being around who she thinks is racist.[3] We would say that in this case it is true that Kautar believes that her friend's father is racist not because there are not enough reasons to suspect about her sincerity or her capacity to know her mind; simply, the whole context triggers the intuition that there is a presumption of authority, i.e., her sincerity intuitively guarantees the truth of her self-ascription. In other words, her sincerity is the most salient contextual feature in order to determine the truth of her self-ascription. That's our paradox. These cases, together with the ones introduced below, have the objective to show the high context-sensitivity of mental self-ascriptions.

Recall the desideratum DISANALOGY: A suitable concept of polarization must accommodate the disanalogy between self-ascribing a mental state and expressing a mental state in order to accurately measure polarization. This desideratum was embraced as a result of discussing the problems that self-report questionnaires encounter to measure, in a precise way, the attitudes that they target, along with some strong evidence against the first-person authority thesis. The authority thesis, in its weak and more plausible version, can be defined as follows:

[2] This case is inspired on Nour's case from Srinivasan (Srinivasan 2020).

[3] In order to strengthen the claim that it is intuitively true that Kautar believes what she says she believes we have carried out a survey of 20 participants. Only 4 of them have the opposite intuition, namely that Kautar does not believe what she says she believes.

> **Authority**: There is a presumption according to which the speaker's sincerity guarantees the truth of her mental self-ascription.

Lack of authority, and therefore the inability to accurately identify the state of mind we are in, points in the direction that there is a crucial difference between what we say about the state of mind in which we are and the state of mind in which we actually are, which we express through the things we say and do. But, then, when people answer the tests designed to measure polarization, what kind of information are they conveying? What do we do when we ascribe a belief, or another mental state, to someone or to ourselves? What type of theory concerning mental state ascriptions can accommodate the desideratum DISANALOGY?

We need an approach to the mind, specifically to mental state ascriptions, that accommodates the disanalogy and rejects the authority thesis. However, such an approach also has to accommodate our intuitions toward a variety of cases from different contexts where, sometimes, the speaker's sincerity does guarantees the truth of their mental self-ascriptions. Thus, such an approach should accommodate cases such as CASE 1 and CASE 2.

As a first introduction of what cannot be taken as a right understanding of the mind, take the following quotes: "People use central attitudes all the time; hence they are well established *in the brain*" (Hetherington & Rudolph 2015: 99, our italics); "Whereas research on affective polarization delves into mental processes *inside the voters' heads*, a different line of research examines the physical location of voters' heads" (Fiorina 2017: 63, our italics). Despite these quotes pointing just to a natural way of talking about our mental states, they suggest an intuitive but wrong picture of what the mind is, at least for the purposes of this dissertation. These claims suggest that our mental vocabulary accomplishes a descriptive function, that is, its meaning, and the truth-value of the claims about our minds, depend on the state of affairs that these claims describe. The general idea of descriptivist views, then, is that our mental ascriptions, such as the claim "A believes that p", describe a disposition of objects, either internal or external to the subject, that, when it is the case, it is a fact. Thus, if the state of affairs described with a mental ascription is the case, then the belief ascription is true. For instance, if someone holds that beliefs are equivalent to certain brain states of a subject, then that person would say that if it is true that Oscar's brain is in a certain configuration, then it will be true that Oscar believes what he says he believes. This view is an instance of a more general position, the relational theory of mind, according to which to be in a mental state is to be in a relation with something else, be it a brain state, a proposition, or whatever.

In this chapter, we discuss different descriptivist approaches to mental ascriptions, and argue that these views cannot satisfy the desideratum DISANALOGY and accommodate at the same time our intuition regarding cases like CASE 1 and CASE 2. That is, different cases of belief self-ascriptions, where the speaker at times exhibits authority and sometimes does not, cannot be homogeneously accommodated by descriptivist approaches to mental self-ascriptions. The structure of the chapter is as follows. In section 5.1, we first delve a little further into the distinction between self-ascribing a mental state and the mental state we actually are in. Then, in section 5.2, we discuss whether a broad family of theories about the mental, namely descriptivists, and two particular contemporary positions –Bar-On's neo-expressivism (section 5.3) and a version of Srinivasan's epistemological externalism (section 5.4), can accommodate not only the gulf that seems to exist between self-ascribing and expressing a mental state, but also the other intuition of the paradox, namely: that someone sincerely and spontaneously asserts that she believes or feels that p, then in certain cases it is highly intuitive to consider her mental self-ascription as true. Finally, we discuss some alleged peculiar types of mental states different from standard beliefs, the so-called *aliefs* (Gendler 2008), *unendorsed beliefs* (Borgoni 2018b) and *in-between cases* (Schwitzgebel 2001, 2010, 2013), and argue that these concepts do not really point to peculiar types of mental states, but to particular situations pressing us to evaluate whether someone believes or not that p (section 5.5), which unveil the normative nature of the issue of mental ascriptions.

## 5.1. Self-ascribing a mental state Vs. expressing a mental state

Our intuitions regarding the truth of what someone sincerely says about her mind significantly vary from case to case. In some cases, we have no doubt that there is at least a presumption of authority. In other cases, however, what someone says is not enough to determine the truth or falsity of her mental attribution, and not necessarily because the speaker is suspected of being a liar. There are even situations, more complicated but no less common, in which our intuitions are divided and there is no matter of the fact that settles the dispute over whether someone believes that p. To see some other examples of it, consider the following real cases:

1. A group of researchers concerned with the scourge of gender abuse and sexism investigate the brains of abusers and find certain differences. They publish a paper suggesting that gender abuse, and with it at least certain types of sexism, have to do with atypical functioning of the brain.

2. Inter Milan fans, after uttering racist insults toward the football player Romelu Lukaku during a match, wrote a letter to Lukaku saying that they are not racist at all, and that these chants were a form of respect: "We are really sorry you thought that what happened in Cagliari was racist. You have to understand that Italy is not like many other north European countries where racism is a REAL problem".

3. Many people in the USA think they dislike Elizabeth Warren just because of her controversial ancestral claims and her progressive economic views. However, when similar views are held by a man who makes similar or even more controversial claims, the citizens feel less aversion toward him.

4. Many citizens of Ceuta, a small Spanish city characterized by its high cultural diversity, sincerely say that Ceuta is an example of a diverse society, where different cultures live together in harmony and peace. However,

some of them perceive with great rejection and aversion the possibility that their son or daughter gets married to a person from another culture.

5. A New York Times columnist, Bret Stephens, received a joke on Twitter from an Associate professor at George Washington University comparing him to a bedbug. The joke had little impact (9 likes, 0 rts), and Stephens was not explicitly mentioned in the tweet. Stephens emailed to the professor complaining about his joke, with a copy to the provost of the university where the professor works. The professor made public Stephens' email, and Stephens ended up having to deactivate his Twitter account due to complaints. In a later interview, Stephens said that he had not intended to cause the professor any kind of professional trouble. However, given that he sent a copy to the provost and once he compared the professor's harmless joke with the rhetoric of totalitarian regimes, the professor claimed that Stephens did intend to cause a problem to him: "Cc'ing the Provost meant that he was trying to use his social status to get me in trouble. And that means it isn't about civility at all; it's about power".

6. A woman sincerely confessed a crime she hadn't committed: she sincerely said she was guilty, despite having claimed for some time that: "in my head and in my heart, I know I wasn't there". Despite the fact that many pieces of information from the case didn't fit with her confession, the jury thought that no one can know better than oneself what one knows. Thus, the woman entered into prison. She was eventually exonerated by DNA evidence.

All these cases involve, in one way or another, a sincere mental self-ascription. Your intuitions, we suppose, about the truth of what they say about themselves, and about whether their sincerity is enough to guarantee the truth, shake in some cases, and are clearer in others, pressing in both directions. On the one hand, it seems that it makes sense to think that nobody knows better than oneself what one believes, as Finkelstein (Finkelstein 2003: 9) and many others pointed out, and as it was assumed by the jury in the latter case, or as we sometimes assume in our daily linguistic exchanges. On the other hand, it seems that the rest of cases, to a greater or lesser extent, trigger an opposite intuition. For instance, it is quite clear that Inter fans, given the things they claimed, believe that Lukaku belongs to an inferior group of people because of his physical appearance, despite explicitly saying that they are not racist. And it seems that many USA citizens dislike Warren because she is a woman, despite their explicitly saying that it is because of her policies and previous controversial statements. In that sense, it seems that the speaker's sincerity does not guarantee the truth of mental self-ascriptions in some cases. Mental self-ascriptions seem to be highly context-sensitive.

As advanced in section 3.4.1, there are some reasons to be suspicious about the authority thesis, even in its weak version. In particular, we contend here that, with some regularity, there is a deep gulf between what we sincerely say we believe and what we really believe –not only because of self-deception or confabulation cases, and we are very often wrong about which side of the dividing line we are on (Hurlburt & Schwitzgebel 2007; Schwitzgebel 2011a,b). Take the following MacFarlane's quote:

> It is crucial to mark the distinction between *expressing* one's liking for a food and *asserting* that one likes the food. One does the former, but not the latter, when one smacks one's lips in delight after a good meal. One does the latter, but perhaps not the former, when one tells one's host, with an unconcealed expression of dutiful weariness, that one liked her cooking. (MacFarlane 2014: 15, our emphasis)

Asserting that one likes the food is not necessarily the same as actually liking it, even if it is sincerely asserted. Usually, sincerely and spontaneously asserting that one likes the food counts as a way of expressing that one likes the food because it normally functions as an evaluation (see chapter 7). But one can end up discovering that one does not actually like the food despite having asserted such a thing. In this case, the assertion wasn't a genuine evaluation. On the contrary, if one expresses that one likes the food (by smacking one's lips in delight after a good meal, but also by making some genuine positive evaluations about the meal), then one likes the food, no matter the mental self-report one does (see chapters 6 and 7). Since asserting that one is in a particular mental state is not necessarily the same as being in that particular mental state, especially when talking about our mental states towards complex topics, it is important to pay attention to what people express rather than to what people self-report in order to measure polarization.

## 5.2. Descriptivist approaches to mental ascriptions

Here we will be concerned with descriptivist approaches to mental state self-ascriptions. But before that, we would like to briefly distinguish some orthogonal theses that sometimes are taken to be analogous to this issue, or to be part and parcel of it. The first thesis is the core one of the representational theory of mind, quite widespread in the cognitive sciences. According to this theory, our thinking is the processing of *mental representations*, and mental representations are *physical things* with meaning (see, for instance, Shea 2018: 4). Our mind, then, basically consists in our capacity to represent the world and operate with those representations, which are physical things that represent objects and properties in the world –their contents. From this approach, mental representations are inner things located somewhere inside our heads. Brentano called this strange way of existing "intentional inexistence" (Brentano 1874).

In philosophy and psychology, the study of the mind has been mostly related to ontological (e.g., what is the mind?) and epistemological (e.g., how can we know our mind?) concerns. Representational theory of mind usually has a foot in both fields: in ontological terms the mind is taken to be a representational system, and in epistemological terms the theory holds that we know the world and our own mind through the representations we form –of course, one can be a representationalist because of being committed just to one of those theses.

Cartesianism, which is a set of theoretical commitments that can be traced back to Descartes' philosophy and that includes the representational theory of mind, is still highly influential in the philosophy of mind and the contemporary cognitive sciences. For Descartes, the ontological commitment behind the representational theory of mind is supported by the stance that mind and body are two distinct substances, and the epistemological commitment rests on our privileged and direct access to our mental life through introspection. But there is a third commitment, more subtle, that is typical of Cartesianism: semantic descriptivism, or just descriptivism. Behind the ontological and epistemological dimensions of Cartesianism it is the semantic commitment that the function of our psychological language is to describe a portion of the world, i.e., to refer to those facts that compose our mind, describable in terms of objects and first-order predicates. Although many contemporary positions have abandoned the ontological and epistemological dimensions of Cartesianism, the semantic commitment is still quite pervasive (Pinedo 2020; see also Nuñez de Prado-Gordillo in press, especially chapter 2, for a recent discussion on it), and has been a shared feature by prominent positions in the discipline, such as the mind-brain identity theory, functionalism, emergentism, eliminativism.[4] Thus, according to descriptivism, the function of a mental self-ascription is to describe or represent a possible or real state of affairs, i.e., a specific distribution of objects that, when is the case, is deemed a 'fact'. In this section we will focus on those positions that hold that mental self-ascriptions describe states of affairs, facts, that determine their truth-value.

One trend within mental descriptivism holds that the facts making a mental state ascription true are internal to the bearer of the mental state (see, for instance, Clutton 2018). According to these theories, if I sincerely self-ascribe the belief that my phone is on the table, the sentence "I believe that my phone is on the table" plays the function of describing a state of affairs, i.e., it refers to a fact somehow taking place inside my head, specifically in my conscious experience, which makes my self-ascription true. If that state of affairs is not the case and I am not internally in a relation to such a proposition, then my self-ascription is false. Following Finkelstein, we call these 'detectivist' positions. As Finkelstein

[4] In general, all the positions within the framework of what is called 'the standard view' (see Sánchez-Curry 2018), according to which beliefs are theoretical cognitive entities that play a role in our cognitive systems, are descriptivists. However, some versions of certain positions that criticize the standard view, such as certain versions of dispositionalism and interpretativism, are still committed with descriptivism insofar as they assume that the function of belief ascriptions is to describe something, and hence the truth-value of a belief ascription depends on, for example, that the agent fits certain dispositional properties of a certain dispositional stereotype or certain social norms that are given once and for all.

puts it,

> A detectivist is someone who believes that a person's ability to speak about her own states of mind as easily, accurately, and authoritatively as she does may be explained by appeal to a process by which she finds out about them. According to detectivism, I am able to state my own thoughts and feelings because they are conscious, and they're conscious thanks to a cognitive process by which I have detected their presence". (Finkelstein 2003: 9)

Thus, the facts referred to by mental self-attributions, and which confer truth to the latter, are internal to the subject, who detects them.

But a descriptivist regarding the mental might hold that the facts making true a mental self-ascription are external to the speaker's body, being social or otherwise (see Macdonald 1995: 99; Villanueva 2014: 54). Thus, according to another trend within descriptivism, what makes a mental self-ascription true is an external fact: the truth-value of a mental attribution is supported by empirical evidence. A strong point of externalist descriptivism is that it denies the Cartesian assumptions of what Ryle calls "the official doctrine" and "the dogma of the Ghost in the Machine" (Ryle 2009: 1-8), i.e., the idea that there is an 'inner world', in contrast to the external world, that somehow exists inside our bodies, and that we know better than anyone. Thus, according to externalist descriptivism, if I sincerely self-ascribe the belief that there will be a resurgence of coronavirus, the sentence "I believe that there will be a resurgence of coronavirus" plays a descriptive function; but instead of describing something internal, it refers to some external empirically observable facts, which determine its truth-value. For instance, if we are behaviorists, certain pattern of behavior would suffice to claim that I actually have the belief I self-ascribe. Or if we are social externalists regarding the mental content (Burge 1979; see also Baker 2007 and Frápolli & Romero 2003), the fact that I master the use of the concept of coronavirus in my community might suffice to claim that I actually believe what I say I believe, while in another different community where the word 'coronavirus' has another meaning my sincere self-ascription would be false. But typically, many philosophers of mind, especially reductionists and eliminativists, take our brain states as the relevant external facts, which, despite being located inside our body, can be approached from an external perspective.

The first thing to note is that detectivism does not seem to be in a good position to deal with the concerns discussed in chapter 3 regarding how to accurately measure polarization. In particular, it does not allow us to explore indirect ways of measuring people's attitudes, because it states that people are always or commonly authoritative, given that the facts making true a mental ascription are internal to the subjects. In that sense, detectivism is incompatible with DISANALOGY. To see it more clearly, consider CASE 1 and CASE 2. Detectivism can accommodate our intuition in CASE 2, for example by stating that Kautar's doxastic self-ascription is true because the fact responsible for it is internal to her, i.e., it is inside Kautar, and in this respect, her self-ascription will be normally true no matter what happens outside her (except, perhaps, in some exceptional cases). However, detectivism cannot accommodate our intuition in CASE 1: according to detectivism, Oscar's self-ascription should be considered as true because mental self-ascriptions are always authoritative. One may think that CASE 1 is one of the exceptions that can be assumed by detectivism. But if this case counts as exceptional, then a large number of cases of mental self-attribution would, and then there would be no general presumption of authority in mental self-ascriptions.

Externalist descriptivism, on the other hand, seems to be better positioned than detectivism to deal with the concerns of how to accurately measure polarization. Since externalist descriptivism states that the truth-value of a mental self-ascription depends on external facts, an obvious strategy for an externalist to pursue is to argue that, despite appearances, people are not in fact authoritative about their own mental states, and then it follows that there is no presumption of authority regarding our own mental states (see, for instance, Boghossian 1989), at least in the non-epistemic sense we are discussing here.[5] In that sense, it can ac-

[5] Two things. Of course, there are externalist positions holding first-person authority. We will discuss later a version of these positions that we attribute to Srinivasan (section 5.4). However, it is important to note that many externalist views about the mind aren't externalists regarding the truth of a mental self-ascription, but regarding the truth of the *content* of the mental self-ascription,

commodate DISANALOGY: it acknowledges that there can be a gap between what we say we believe and the attitudes that we hold, because what determines the truth of a mental self-ascription is an external fact, and therefore the speaker's sincerity does not commonly guarantee the truth of her mental self-ascription. However, it is not easy to imagine how the idea that the truth of a mental ascription depends on an external fact, or more generally the idea that having certain mental states is partially a matter of external facts, would materialize in order to measure polarization. Would we need to discover the facts on which it depends that someone really has a certain belief? Additionally, this position, like detectivism, cannot accommodate our intuitions regarding CASE 1 and CASE 2, that is, it cannot take into account, homogeneously, different cases of mental self-ascription where the speaker sometimes exhibits authority and sometimes does not. Contrary to detectivism, externalist descriptivism could accommodate our intuition in CASE 1, for example by saying that it is false that Oscar believes what he says he believes because external facts do not support the truth of his self-attribution. However, regarding CASE 2, this position would also have to say that it is false that Kautar believes what she says she believes, because there is no

that is, they are about how contents of thought are individuated. In this sense, many of them are compatible with first-person authority. Consider my thought that the washing machine is on, a first-order thought directed toward a state of affairs beyond my body. Some externalists would say that I am not authoritative about this thought: the thought could be false, and I am in no better a position than anyone else to know whether it is, since whether it is depends on how things are in my environment. But now consider my thought that I am presently consciously thinking that the washing machine is on. Many of those externalists would say that I am authoritative about this thought (Davidson 1987). This thought is not only directed toward another state of mine, with which it is co-present; it contains as part of it that first-order thought. In thinking that I am thinking that the washing machine is on, I am thinking that the washing machine is on. Since this last first-order thought of thinking that the washing machine is on depends on how things are, one can be an externalist and still holds first-person authority: I have authority regarding my mental states and their truth depends on the world. That's how Burge and Davidson argue in favor of the compatibility between externalism and first-person authority (see Macdonald 1995: 104). But again, note that this position is externalist regarding the content of a mental state but not regarding the truth-value of the mental self-ascription itself. Then, the content of a belief might be true and still be false that the person who self-ascribe such a belief actually believes it, and that's the reason why they are compatible. Much of this is behind the discussions about externalism and self-knowledge.

fact that supports the truth of her self-attribution, and then it remains the case that there is not first-person authority. At best, externalist descriptivism could claim that the relevant facts to determine whether Kautar's self-attribution is true have not been discovered yet, forcing us to suspend our judgment. But CASE 2, like many other cases, triggers the intuition that, sometimes, a speaker's sincere mental self-attribution is true. If this position, at best, forces us to suspend our judgment, then it neither accommodates our intuitions nor offers an acceptable political recommendation.

We would like to end this section by discussing a possible objection, namely the possibility for descriptivists to adopt a kind of hybrid descriptive theory. In particular, someone might think that insofar as a descriptivist theory can defend that in some contexts the relevant fact to determine the truth of a mental self-attribution is internal, and in other contexts this fact must be sought outside, in the world, then such hybrid position could accommodate our intuitions regarding CASE 1 and CASE 2, and could also satisfy DISANALOGY. As Villanueva has already pointed out, such liberal descriptivism based on a certain degree of contextual flexibility has at least two problems (Villanueva 2014: 63). The first of them is that this theory must provide theoretical resources allowing us to sharply distinguish the contexts in which the relevant fact will be internal from those in which the relevant fact is external. The second is that, consequently, the theory must provide two different semantic strategies to give the truth-conditions of mental self-attributions. Instead of delving into these two objections, which we find hard to deal with for descriptivism, we would like to briefly complement the critique toward such hybrid descriptivism by focusing on three points which are important for the objectives of this dissertation.

The first one has already been advanced, and it has to do with how useless the descriptive strategy is in offering recommendations regarding how to measure polarization. We would have to be able to first distinguish internal contexts and external contexts to design different strategies to measure the mental states of the population. And, even if this could be done, the contexts in which the relevant fact is internal to the subject would still be exposed to the problems discussed in section 3.4.1.

The second problem is that descriptivism of the mental, however flexible it may be, entails a kind of eternalism: once the fact that makes a mental self-attribution true is found, that mental self-attribution will be true eternally.[6] And if one rejects eternalism and remains open to a kind of revisionism according to which we can always find a fact to replace the previous one, then this suggests that we should suspend our judgment until science has completed its investigation, in a similar line of the argument adduced by Fodor against semantic externalism (see Fodor 1979). Whichever option is adopted, it seems incompatible with our intuitions: mental ascriptions are not eternally true, and we don't think that a mental ascription cannot be known to be true nor false until we discover all the relevant facts.

Finally, we would like to note that this flexible descriptivism cannot explain how it is possible for the truth of a mental self-attribution to vary if the set of possibilities that we take into consideration changes (Lewis 1996), nor can it account for our intuitions regarding certain situations of disagreement (see MacFarlane 2014). If the truth of a mental attribution depends on internal or external facts, then any situation of disagreement with respect to whether someone believes that p can be settled by appealing to facts, i.e., it will necessarily be a factual disagreement. However, this idea seems to violate our intuition in many cases where the disagreement involved appears to be not-straightforwardly factual (Field 2009). We will return to these questions later (chapter 6).

## 5.3. Bar-On's account

In this section, we deal with neo-expressivism, a recent and influential position on mental self-attributions proposed by Dorit Bar-On, which defends a special security of the first person without necessarily being a detectivist position. This proposal is halfway between descriptivist positions and a family of theories antagonistic to descriptivism: expressivism, whose central tenet is that certain regions of language (in this case, the main use of sentences of mental self-

[6] This idea is different to eternalism in the philosophy of language, i.e., the idea that the parameter time is an element of the proposition expressed by uttering a sentence (Richard 1981).

attribution) do not play a descriptive function, but an expressive one. Concerning this point, Bar-On bets, she says, for saving the difference between descriptive and evaluative functions of language without sacrificing logico-semantic continuities, among other theoretical advantages (Bar-On 2019: 11).

To understand Bar-On's position well, the first thing to note is that her 'expressivist' position departs from the position commonly associated with Wittgenstein's ideas on mental self-attributions, which she calls 'simple expressivism'. According to simple expressivism, i) mental self-ascriptions only serve to express, and in no way report or describe, a state of mind, ii) they are on a par, semantically and epistemically, with nonverbal expressive behaviors such as winces and moans, and iii) they are neither truth-evaluable nor epistemically assessable (Bar-On 2019: 18-19).

The neo-expressivism proposed by Bar-On, on the contrary, maintains that mental self-attributions have an expressive and descriptive function, have semantic content, and can be declared true or false. She argues that, as *acts*, mental self-ascriptions are spontaneous expressions of our mental states. In this respect, as acts, they are interchangeable with verbal expressions like "what a great movie!", and are importantly continuous with non-verbal expressive acts like hugs (Bar-On 2019: 19). For instance, sincerely saying "I love this movie", spontaneously saying "what a great movie!", and spontaneously giving a hug to your friend after watching a movie can all be acts of expressing the same mental state. However, as *products*, mental self-ascriptions employ linguistic vehicles to ascribe to the avower the mental state avowed. The expressive vehicles are truth-evaluable sentences that express propositions about oneself. That is, despite Bar-On's appeals to the expressive character of mental self-ascriptions as acts to explain its distinctive security, she departs from simple expressivism "in highlighting the fact that, like various mental and non-mental descriptive reports, avowals use expressive vehicles –sentence- or thought-token– that are semantically complex and are truth-evaluable" (Bar-On 2019: 20). Thus, she distinguishes between the acts of expressing and the expressive vehicles used, to which she sometimes refers as the distinction between *a-expressing* (as act) and *s-expressing* (as product) (Bar-On 2019: 23), and applies it also to ethical discourse (Bar-On & Chrisman 2009).

Despite the fact that in some earlier works Bar-On (Bar-On 2004, 2015) seems to assume the idea that the truth-value of mental self-ascriptions hinge on some factual matters –inasmuch as she defends a hybrid position designed to keep some benefits from descriptivism– in a recent work (Bar-On 2019) she explicitly rejects this idea by arguing that her proposal does not necessarily entail mental internalism (although the vocabulary she often uses, like the verb 'give vent', suggests that she is thinking of *something* that we *release* through an utterance). To argue so, she adopts a Davidsonian theory of truth, according to which, "s is true if and only if p, where p is replaced by any sentence that is true if and only if s is". Maybe, with this move Bar-On avoids a criticism according to which it counts as a descriptive one because what makes true an avowal is a matter of fact (see for this discussion Villanueva 2014). However, neo-expressivism still cannot avoid a broader objection, which places it in a bad position regarding our goal in this dissertation.

When you make a mental self-ascription as an act, you are sincerely and spontaneously giving vent to your mental state, and that's what determines its special security, its immunity from correction. In other words, a-expressing your mental state guarantees that you are in the mental state you claim to be in, because you are giving vent to it. Thus, mental self-ascriptions, as acts, determine the contexts in which the mental self-ascription can be true: if I sincerely say that I'm in pain, the act of expressing my mental state itself guarantees its truth, because it is its condition as act that explains its special security, according to Bar-On (Bar-On 2015: 141; Bar-On 2019: 20). Therefore, it follows that the contexts in which I avow that I'm in pain are the same contexts in which my avowal is true. Furthermore, if I say that I'm in pain in a non-sincere way, then my self-ascription is automatically false. Let's put it another way: there is no context in which I could sincerely and spontaneously say that I'm in a mental state –i.e., making an avowal– and it nonetheless be false that I'm in that mental state. Avowals have a special security because as acts they "give vent to the very states of mind that the avowals understood as products (that is, qua linguistic or mental representational tokens) ascribe to the avower" (Bar-On 2019: 19). Hence, Bar-On equates truth-conditions with felicity conditions, and this move prevents the possibility of accommodating the distinction between sincerely self-ascribing a mental state and expressing the mental state in which one really is. That is to say, it follows from neo-expressivism that in CASE 1 either Oscar's belief self-ascription is true, or it makes no sense because it does not satisfy the felicity conditions of that claim. Both options contradict our intuition. In this sense, neo-expressivism is contrary to the idea that someone can sincerely self-attribute a state of mind and yet be false that such a person is in the state of mind she claims to be. Therefore, it cannot accommodate DISANALOGY.

## 5.4. Srinivasan's account

Srinivasan has recently defended a *radical epistemic externalism* (Srinivasan 2015, 2016, 2020), a particular sort of "hard-nosed epistemic externalism", as she once called it (Srinivasan 2016: 378). According to this position, the epistemic justification to believe something partly comes from external facts, for example, as she puts it, "whether one's belief exhibits an appropriate causal connection to its content, or is a product of a reliable or safe method" (Srinivasan 2020: 401). In other words: one is epistemically justified in believing that p if there is an external fact that supports such a justification. Despite the fact that this position is strictly speaking about epistemic justification, and not about how the truth of a mental self-ascription is determined, we think that Srinivasan can be attributed, with justice, an analogous position about the truth of self-attributions of mental states, or, at least, about the truth of knowledge self-ascriptions. The reason is that Srinivasan claims that a person who says "I know that p" and has a privileged access to the external fact supporting her epistemic justification, will still know that p no matter how high the stakes become, as in the case of a woman's testimony of rape (Srinivasan 2016: 378). In that sense, Srinivasan is committed to the idea that external epistemic justification guarantees the truth of a self-attribution of knowledge. Thus, according to this analogous position, the truth of a mental self-ascription depends on an external fact. Although at first glance it may seem so, Srinivasan's position is not exactly like the externalist descriptivism discussed in section 5.2. Let's see why.

Srinivasan is first committed with a kind of anti-Cartesianism, based on its epistemological dimension, and defined as the position according to which there are no transparent conditions (Srinivasan 2015: 274). A transparent condition is defined as one in which (i) whenever one is in a state of mind M (belief, desire, feeling, etc.), one is in a position to know that one is in M, and (ii) whenever one is not in a mental state M, one is in a position to know that one is not in M (Srinivasan 2015: 274). Thus, Srinivasan holds a position according to which we are systematically wrong about which are our own mental states. Nevertheless, Srinivasan clarifies that embracing this type of anti-Cartesianism does not imply the skeptical verdict that one is never in a position to know if one is in a particular state of mind (Srinivasan 2015: 275). Anti-Cartesianism is compatible with what Srinivasan calls *contextual transparency*: in certain contexts, one can know if one is in a particular mental state (Srinivasan 2015: 276). This is a good start.

So far, we subscribe every word. Anti-Cartesianism seems to assume that there is no presumption of first-person authority, and contextual transparency allows us to argue that sometimes there is indeed authority. Thus, it looks like this position could accommodate DISANALOGY and our intuitions in CASE 1 and CASE 2. But what determines the truth of a mental self-attribution then?

Regardless of whether we can sometimes be in a position to know whether we are in a particular state of mind or not, Srinivasan argues that what makes a mental self-attribution true (epistemically justified) is an external fact. However, one of the interesting points of Srinivasan's proposal is that she argues that there are cases of *bad ideology*, that is, cases in which subjects live in conditions under which certain "pervasively false beliefs have the function of sustaining (and are in turn sustained by) systems of social oppression: patriarchy, racism, classism" (Srinivasan 2020: 409). In those cases, the truth (epistemic justification) of people's mental self-ascriptions has to do with whether their capacity to self-ascribe a mental state (track the truth) is distorted by ideological forces or not, or whether they are endowed with capacities allowing them to pierce through ideological distortion or not (Srinivasan 2020: 409). In other words: someone could self-attribute an attitude which she might not actually have because she is under the influence of bad ideology. For example, Srinivasan argues that a woman can say that she knows she deserves to be beaten by her husband and yet it is not true that she knows it (for one, the content of her belief is false and it is not epistemically justified), and the reason is that she is a victim of bad ideology (Srinivasan 2020: 410). And, on the other hand, a person can self-attribute an attitude that she actually has simply because, given her social position, she is better placed to know the fact that makes it true.

Thus put, Srinivasan's position seems to be able to accommodate CASE 1 and CASE 2 and satisfy DISANALOGY. In CASE 1, Oscar does not believe what he says he believes because there is no external fact that supports his self-attribution (there is no fact that epistemically justifies his claim). In CASE 2, Kautar does believe what she says she believes because, given her social position, she has privileged access to the reality that guarantees the truth of her self-attribution. This proposal has at least three advantages over Bar-On's position. The first one is that the truth of a mental self-attribution does not depend on someone sincerely self-attributing it. The second is that it does not equate truth-conditions with conditions of felicity. Third, the political goal of the position implies an extra argument in its favor: it allows us to defend that a socially oppressed person knows despite not being believed or suffering gaslighting, which offers a way to fight against epistemic injustice.

We fully agree with the political goal pursued by this position, and we believe that the appropriate theory will be one that allows us to fight against injustice. However, we also believe that this position has some theoretical and practical drawbacks that are undesirable and avoidable. The first one is that, insofar as it links the truth of a self-attribution with a fact, it is one of a descriptive nature. In that sense, this position is subject to the same limitations indicated in section 5.2. Secondly, this descriptive externalism cannot accommodate our intuition that, in some cases, the social position of the person who self-ascribes a mental state, i.e., her allegedly privilege access to pierce through bad ideology and penetrate the moral reality, is not sufficient to guarantee the truth of her mental self-ascription. Third, this position assumes a moral realism with which we are not very comfortable, both theoretically and practically. Recall that Srinivasan assumes that there is a moral reality, that is, moral facts, that makes a mental self-attribution like "I know I deserve to be beaten by my husband" false and that always makes a mental self-attribution like "I know that my friend's father is racist even though I have no reason to explain why" true if uttered by a person of Arab descent.[7]

We need a position that avoids at least most of the limitations that Srinivasan's descriptivist externalism faces and at the same time preserves its strengths, especially its political motivation. In particular, we need a position that allows us to satisfy DISANALOGY, accommodate our intuitions in CASE 1 and CASE 2 recognizing that sometimes there is a presumption of authority, offer the possibility of measuring the mental states people express to be in, and that finally allows us to adopt the political stance of saying, for example, that it is true that a woman knows that she has been sexually abused if she says that she knows it, no matter how high the stakes become.

## 5.5. Types of mental states? Aliefs, In-between, and unendorsed belief notions

So far, we have seen that, for the main purpose of this dissertation, we need a conception of the mental that allows us to distinguish between what people say that they believe and what people really believe, and that also allows us to track the mental states that people actually are in. Furthermore, such a view should accommodate our intuition that sometimes there is a presumption of authority, as well as that in certain particular cases we can say that it is true that someone has the self-attributed mental state, no matter the context, i.e., cases where the

[7] From a theoretical point of view, postulating the existence of moral properties and facts beyond our conceptual practices is controversial, mainly due to its spooky ontological nature, but also because of the difficulty for the theory to accommodate our intuitions as competent speakers about what is right and wrong in different moral cases, and to accommodate certain situations of moral disagreement and mental attribution where there is no fact that can settle the dispute. From a practical point of view, postulating the existence of such a moral reality would suggest that what is right and wrong is given once and for all, and we cannot be sure about what is right and wrong until we discover the ultimate moral facts, and therefore we should suspend our judgment. This has really pernicious consequences. Furthermore, this idea can hardly account for moral progress (see Pérez-Navarro 2019: 170).

person's sincerity seems to guarantee the truth of her mental self-ascription.

In line with this second set of requirements for a theory of the mental, in this section we introduce some cases that have been frequently presented as a particular type of mental state because of the problem they pose. In particular, we are dealing here with extreme cases where our intuitions are unclear with respect to whether a person believes, wants, fears, expects, etc., that p. These cases have been called "in-between believing" and "in-between cases" (Schwitzgebel 2001, 2010, 2013), and recently "unendorsed beliefs" (Borgoni 2018b) when focused on beliefs. An adequate theory of the mental should also account for these cases. Consider the following case from Schwitzgebel:

> **Juliet the implicit racist**: Many Caucasians in academia profess that all races are of equal intelligence. Juliet, let's suppose, is one such person, a Caucasian-American philosophy professor. She has, perhaps, studied the matter more than most: She has critically examined the literature on racial differences in intelligence, and she finds the case for racial equality compelling. She is prepared to argue coherently, sincerely, and vehemently for equality of intelligence and has argued the point repeatedly in the past. Her egalitarianism in this matter coheres with her overarching liberal stance, according to which the sexes too possess equal intelligence and racial and sexual discrimination are odious. And yet Juliet is systematically racist in most of her spontaneous reactions, her unguarded behavior, and her judgments about particular cases. When she gazes out on class the first day of each term, she can't help but think that some students look brighter than others –and to her, the black students never look bright. When a black student makes an insightful comment or submits an excellent essay, she feels more surprise than she would were a white or Asian student to do so, even though her black students make insightful comments and submit excellent essays at the same rate as do the others. This bias affects her grading and the way she guides

> class discussion. She is similarly biased against black non-students. When Juliet is on the hiring committee for a new office manager, it won't seem to her that the black applicants are the most intellectually capable, even if they are; or if she does become convinced of the intelligence of a black applicant, it will have taken more evidence than if the applicant had been white. When she converses with a custodian or cashier, she expects less wit if the person is black. And so on. Juliet could even be perfectly aware of these facts about herself; she could aspire to reform; self-deception could be largely absent. We can imagine that sometimes Juliet deliberately strives to overcome her bias in particular cases. She sometimes tries to interpret black students' comments especially generously. But it's impossible to constantly maintain such self-conscious vigilance, and of course patronizing condescension, which her well-intentioned efforts sometimes become, itself reflects apparent implicit assumptions about intelligence. (Schwitzgebel 2010: 532)

This case is similar to CASE 1, but it is a more limiting case. Does Juliet believe that all races are intellectually equal? According to Schwitzgebel, Juliet's case is a typical in-between case, in which the correct answer is: kind of. "She doesn't fit neatly into the yes or the no, so if we're concerned to describe her precisely, a yes or no won't do" (Schwitzgebel 2010: 537). Schwitzgebel argues that this kind of case poses a problem to representational accounts of mental states (Schwitzgebel 2013: 86), and arguably, it poses the same kind of problem to descriptivist views. The reason is that if having a belief is a matter of fact, then it is hard to explain these kinds of cases where the answer to the question about whether someone has a belief is neither yes nor no.[8]

[8] This claim, while right, is quite rushed. Someone may object that a mental self-ascription can refer to a set of facts and, when only some of those facts hold, the situation counts as an in-between case, because in such a case it is hard to say whether the speaker is in the mental state she says to be in insofar as the whole pack of facts determining the mental state does not hold. Thus, such a position would account for in-between cases despite being a descriptivist one. However, note that this position requires having an exhaustive list of what amounts to believe that p in a particular

These types of cases have received considerable attention in recent years, and the proposals that try to explain them are quite diverse (see, for instance, Bayne & Hattiangadi 2013). For example, based on Gendler's diagnosis of a case where some people display signs of both believing that a transparent walkway over the Grand Canyon is and is not safe (Gendler 2008), Gendler would presumably say that despite the fact that Juliet believes that all races are intellectually equal, she cannot dislodge an attitude that seems to control her behavior as the belief that all races are not intellectually equal would, and that this attitude belongs to a distinct category of mental states, that of so-called *aliefs* (Gendler 2008), which are characterized by being resistant to rational revision. In a similar line, Razinsky have recently defended that ambivalence, defined as a situation where a person holds "two opposed mental attitudes toward one and the same object" (Razinsky 2017: 16), does not point to an irrational or contradictory phenomenon, but just to our common way of being: "The main thought behind this book is that if human lives are in fact often ambivalent, this may be conceived as an invitation to rethink our notions of personhood and rationality, as well as those of mental attitude, desire, judgment, emotion, action, and consciousness" (Razinsky 2017: 4). In saying so, Razinsky assumes that we often hold two opposed mental states toward the same thing. Furthermore, Borgoni has defended that in cases like Juliet's, the subject has two mutually contradictory beliefs (Borgoni 2015b,a, 2016), and more recently has argued that these cases involve a single belief with very peculiar psychological aspects (Borgoni 2018b). Rather than considering them in-between cases or cases involving two contradictory beliefs, Borgoni argues that these are cases of *unendorsed beliefs*. These beliefs, which are deemed as a particular type, are characterized by three features: they remain in our psychology and guide our action despite our explicit rejection of their content, are resistant to our control, and are very hard to know from the first person perspective (Borgoni 2018b: 49). Thus, according to Borgoni, Juliet has the unendorsed belief that all races are not

context, and for such a list, whatever it is, it is not difficult to find a case where, although all the facts hold, it is intuitive to say that the person does not believe that p. Besides, the idea that such a list can be established is doomed from the beginning: the question of what amounts to believing that p is not a matter of facts, but a normative one.

intellectually equal.[9] Juliet's racist belief guides her behavior, is resistant to her control, and is not always aware of it from her perspective. In that sense, Juliet has an unendorsed belief.

We agree with Borgoni that this case, rather than involving an in-between mental state, an alief, or two opposed mental states, is about a single belief, but we don't think that this is a peculiar type of belief. In fact, our intuition in this case is that Juliet does not believe that all races are equal and, since she is aware that she does not believe so, she strives to change it. To the extent that one can predict the behavior of someone who reacts in a racist way (Juliet's case), and the attribution of beliefs has among other aims that of predicting and explaining behavior, it seems that there is a belief that we can attribute to the subjects of the supposed in between cases. Thus, in opposition to Borgoni, we think that this case does not show that there are *beliefs of a peculiar type* due to the particular traits they exhibit. Rather, we think that these *cases* are peculiar simply because they show that the truth of claiming that someone is in a certain state of mind is both highly context-sensitive and a normative issue, which coheres with the idea that sometimes we feel that some self-ascriptions are true simply because the speakers look sincere to us, sometimes we think they are false, and finally sometimes we have unclear or divided intuitions. Possibly, the richer the context, the more difficult will be to determine whether or not someone believes that p. And our intuition will possibly vary when modifying the context. So, according to the explanation we prefer, Juliet's case, then, unveils that the nature of mental attributions is normative, where being normative means that the truth-value of a mental attribution does not hinge on a matter of fact but rather is relative to certain normative practices, together with the possibility of error in claiming that a person, or oneself, is in a particular mental state. In other words: according to our preferred view, to attribute a mental state is to *evaluate*, that is, to make a decision, to have chosen what the relevant features to determine whether someone has a certain mental state are, which in turn expresses our own commitments. But the perspective on which we rely to make our evaluations are not purely subjective:

[9] Endorsing a belief, as Borgoni understands it, is to consciously and sincerely assert or mentally affirm the belief's content (Borgoni 2018b: 55).

they are tied to public rules, which are linked to our social practices, to our way of living. In-between and other similar cases seem to be more plausible when approached from a pure theoretical context. But, if the same issue is approached from a practical context where we have to decide whether a particular subject believes or not that p, the normative character of our mental ascriptions makes more explicit: these cases force us to make a choice, to make an evaluation.

However, this rough sketch of our preferred position does not detract from the task carried out by Borgoni, Schwitzgebel and many others. In other words, we think that it is very useful and important to try to clarify the features shared by different *situations* in which we say of someone that she believes that p. Nonetheless, we think that when the underlying motivation to do so is to distinguish different types of beliefs in an ontological sense, or to distinguish peculiar mental states, rather than pointing to the features of certain situations, then the task is flawed from the beginning: having a belief has to do with our conceptual commitments linked to certain courses of action, and there is no kind of fact that supports the truth of someone having a certain state of mind. In that sense, we argue, mental attributions are evaluations and, to some extent, objective (insofar as they draw on public criteria).

We devote the next chapter to clarify the view on the mental that we consider the most adequate and, more importantly, the one that enables us to defend that affective polarization cashes out polarization in an accurate way because it targets the attitudes that people *express* to be in.

To sum up what has been discussed in this section: cases like Juliet's give weight to the challenge that a proper approach to the mental must face in order to be acceptable. What does it happen in situations where one part of the information presses in one direction and other part of the salient features of the context presses in the other direction regarding whether someone is in a certain state of mind or not? (This is exactly the situation we encounter with regard to the phenomenon of mixed evidence disagreement). A theory about the mind must provide an answer to this question which, to some extent, poses a problem to the goal of accurately measuring polarization. Thus, in addition to satisfying DISANALOGY and accommodating our intuitions in CASE 1 and CASE 2 by highlighting the context-dependence of the truth of a mental attribution as well as its *objective* character, allowing us to trace the mental state in which someone really is in order to measure polarization, the position on the mental that we are pursuing here should allow us to: a) adopt the political stance that in some cases we can say that someone has a certain mental state no matter how high the stakes become, b) accommodate our intuition that different types of disagreement that may arise regarding whether someone has a mental state, and c) deal with borderline cases like the one discussed in this section.

## 5.6. Conclusion

In this chapter, we have discussed some descriptivist views regarding attitude ascriptions. The main reason for that was that we need a theory of mental ascriptions compatible with an operational notion of polarization, that is, a notion that enables us to measure as accurate as possible, and as soon as possible, the type of polarization that endangers some contemporary democracies. For this, we need to measure people's mental states avoiding to ask them directly, because it could be the case that the mental states that people sincerely self-ascribe are not the mental states in which they actually are. So, we need a theory able to satisfy the desideratum DISANALOGY, but respecting at the same time our intuitions in different cases of mental ascriptions.

We have argued that certain descriptivist approaches to the mental, especially those committed with the first-person authority thesis, as well as Bar-On's account, cannot satisfy the desideratum of DISANALOGY. Srinivasan's account, on the contrary, appears to satisfy it, but, as we have seen, it cannot accommodate well our intuitions in CASE 1 and CASE 2, and faces other problems regarding how to measure people's mental states. So, we need to take into consideration another, nondescriptive approach to the mental, and see whether such an approach can meet the desideratum DISANALOGY without violating our intuitions triggered by some other cases where the speaker exhibits authority. We will devote the next chapter to discuss such a possibility.

# Chapter 6

# A Wittgensteinian Picture of the Mind

> "Mental" for me is not a metaphysical, but a logical, epithet. (Wittgenstein LWPP II: 217)

In 2014, Elliot Rodger committed an act of terrorism toward women in the United States. He murdered at least 2 women, and injured at least 14 more people, after sharing through YouTube and through an extensive writing his misogynistic motivation. According to him, his actions were a reaction to an injustice he was suffering, the injustice of having been forced his entire life to be an involuntary celibate because women had never been attracted to him. This tragic attack gave rise to a sort of movement known as the "Incel Rebellion": Rodger imitators began to appear, who usually ended up taking their own lives after their acts of misogynistic terrorism.

Incels routinely use a dehumanized and objectifying language when speak about women, referring to them as non-human animals or mere sexual objects, among other things. Moreover, they routinely claim that they are at the bottom of an unfair hierarchy of attractiveness, and that they are the real victims. Through these claims, they appear to self-report their beliefs that women are non-human animals, or animals without mind as they sometimes refer to them, and that there exists an unfair hierarchy of attractiveness. But do they really believe such things?

It might be tempting to conclude that they do falsely believe that women are non-human animals, mere mindless things, and that there is an unfair hierarchy that they are victims of. At least, that's what would be suggested by advocates of the first-person authority thesis: to the extent that their claims are sincere, we should say that it is true that incels believe such things. However, it is better to proceed with egg-shells here. As Kate Manne has pointed out in her book *Entitled. How Male Privilege Hurts Women*, to conclude that incels believe what they say they believe concerning these things would be wrong (Manne 2020: 14-32). The reason is that these claims appear to be incompatible with the rest of things that incels say through their videos and writings where they report their motivations. First of all, incels repeatedly express their wish to be desired and admired by women, and wonder why they are not attracted to them. In this sense, incels recognize the mental life of women, as well as their agency, because to wish being desired and admired by women presupposes that they are minded, free agents. Second, given their other concomitant racist beliefs, closely linked to their misogynistic ones, it can be stated that they do not really think that the ”hierarchy of attractiveness” is unfair; in fact, they love hierarchies, especially racial ones. They simply want to be located on the top of those hierarchies because they feel they deserve it. So, although Rodger and others may sincerely self-attribute the beliefs that women are mindless beings and the like, it can be argued that they do not actually believe such things. Insofar as having a mental state depends on the conceptual links established between a mental state and the things a subject says and does, 'mental' is a logical epithet, as Wittgenstein says. So, it can be the case that someone does not actually believe what he says he believes, even though he self-ascribes that belief in a sincere manner.

But there is another different possibility regarding what we do when we self-ascribing a belief. Let's put another example. In 2004, Barack Obama gave a speech at 2004 Democratic National Convention. In his speech, Obama appealed to the idea that citizens of the United States actually think alike at heart. He expressed this idea in different ways: "there's another ingredient in the American saga, a belief that we're all connected as one people", "It is that fundamental belief: I am my brother's keeper, I am my sister's keeper that makes this country work.

It's what allows us to pursue our individual dreams and yet still come together as one American family. E pluribus unum: ”Out of many, one.””. But, in particular, he explicitly made certain belief self-ascriptions in his speech:

> I believe that we can give our middle class relief and provide working families with a road to opportunity. I believe we can provide jobs to the jobless, homes to the homeless, and reclaim young people in cities across America from violence and despair. I believe that we have a righteous wind at our backs and that as we stand on the crossroads of history, we can make the right choices, and meet the challenges that face us. (Eidenmuller 2008)

In this case, his belief self-ascriptions seem to play a particular role beyond reporting some of his beliefs. Of course, in this case we can also discover that Obama didn't believe that we can give our middle class relief and provide working families with a road to opportunity, etc. But these belief self-ascriptions, here, seem to accomplish another different function beyond reporting them. It is not, as it has been said, that “Barack Obama burst onto the national scene with a speech denying the power –denying even the reality– of the deep divisions that seemed to define American politics” (Klein 2017). Through his speech, he is *expressing* certain attitudes especially linked to action. He is expressing his commitment to do certain things, and not just reporting what he believes. Given the context, understood in a broad sense, it can be said that the meaning expressed by Obama's speech is conceptually tied, in an especial way, with certain courses of action beyond those conceptually tied to simply having such beliefs. Thus, through a belief self-ascription one can not only say that one believes such a thing, but also expressing something else, certain affective attitudes, i.e., those attitudes especially tied to action. In other words, by making a belief self-ascription, or even by simply asserting p, one expresses her belief that p, and it can be the case that one does not actually believe that p in virtue of its conceptual connection with certain courses of action. But besides, by making a belief self-ascription, one can also express certain affective attitudes, those attitudes that have an especial link to action. ‘Expressing’ works differently in these situations: one might not have the specific belief that one expresses through one's assertions or belief self-ascriptions, but it cannot be the case that one does not have the affective attitudes that one expresses through one's verbal and nonverbal behavior (see chapter 7).

The recognition of these possible situations is fundamental to the topic of polarization, more specifically to the issue of exactly what should be targeted to measure polarization, and how we should do it. Political polarization often has to do with how public opinion is distributed, which has to do with ascribing beliefs. If belief ascriptions, both in the first and the third person, are taken as accomplishing a descriptive function, then some cases of belief ascriptions cannot be taken into account (see chapter 5). We need an approach to belief ascriptions, and to certain mental state ascriptions in general, that enables us to accommodate all kinds of cases of belief ascriptions, and that accommodates as well the two types of situations of belief self-ascription exemplified with the two cases presented above. One might not have the sincerely self-ascribed belief, but one might also express other practical information through one's belief self-ascription. The first possibility is crucial to measure ideological polarization, that is, to measure exactly the contents believed by a population, and not those they self-ascribe. The second one is crucial to measure affective polarization, that is, to measure the practical attitudes closely connected with having a certain level of radicalism.

In this chapter, we will paint a parsimonious and perhaps unintuitive but operational picture of the mind: a conception that allows us to go one step further in the direction of explaining why affective polarization's tools enable us to measure the type of polarization that endangers democracy, or at least to explain why they are in the right pathway. This chapter, together with chapter 7, should be read as a unit that will provide the theoretical tools necessary to reassess the concept of affective polarization as we do in this dissertation, that is, the theoretical tools assumed by our notion of polarization in attitudes.

According to the picture we offer in this chapter, the mind is not something inside our heads, essentially because it is not a thing, and therefore it has no location. On the contrary, it holds that mental vocabulary belongs to the domain of the normative, instead of the descriptive (Almagro et al. 2022 forthcoming; Fernández & Heras-Escribano 2020; Frápolli 2019; Frápolli & Villanueva 2012; Heras-Escribano

& Pinedo 2016; Pinedo 2020): it has to do with our commitments to follow certain courses of action, which must not be conflated with behaviorism.[1] Our view, contrary to the view according to which we always exhibit a presumptive authority regarding our mental self-ascriptions or we never exhibit such a feature, holds that our mental state ascriptions are dependent on certain norms. In this sense, we claim, we exhibit *contextual authority* (see section 6.6). To ascribe a belief to others or to oneself is to follow certain rules. But, crucially, the rules we think we follow might not be the rules that we actually follow. This approach, then, takes into account the possibility of error when ascribing mental states. More specifically, this view enables us to distinguish two different types of error, mentioned above. It can be the case that we self-ascribe a belief in a sincere way and, actually, we don't have such a belief. But also, it can be the case that we self-ascribe a belief in a sincere way and, through it, we express certain attitudes especially linked to action. These are two very different situations, with very different consequences for measuring polarization.

Our proposal here is to some extent in line with a view recently defended, called *relativistic Rylean view*. According to this view, "to have an attitude of belief is to live −to be disposed to act, react, think, and feel− in a pattern that an actual belief attributor identifies with taking the world to be some way" (Sánchez-Curry 2018: xvii). This relativistic Rylean view takes distance from other interpretativist positions that assume that the norms from which we decide whether someone believes or not that p are given once and for all. Of course, our proposal is also in line with some Ryle's key remarks. In particular, it is in line with (i) the idea that mental expressions such as 'believes that' usually mean that an individual "is prone to do and feel certain things in situations of certain sorts" (Ryle 2009: 116), (ii) the idea that belief ascriptions are inference tickets that license some inferences (e.g., attributing to A the belief that the supermarket is closed is possessing a ticket that licenses the inference that A won't go to the supermarket), and (iii) the idea that attributions of belief should not be construed "as asserting extra matters of fact" (Ryle 2009: 119). However, against certain dispositionalist

[1] Behaviorism is another reductive materialist position that swallows the bullet by approaching the mind in ontological terms plus reducing the psychological domain to a descriptive one.

accounts that take inspiration from Ryle's account, such as Schwitzgebel's dispositionalism (Schwitzgebel 2001, 2010, 2013), we don't think that believing that p is to match to an appropriate degree and in appropriate respects a set of dispositional properties of the dispositional stereotype for believing that p (Schwitzgebel 2002: 251), mainly because this view seems to assume descriptivism regarding the set of properties that compose a dispositional stereotype. Our proposal is normative in nature, and has in its core the idea that it is possible to fail at identifying our own mental states.

The antidescriptivist and pragmatist view we offer here allows us to go a step further in explaining why the notion of polarization that we offer in this dissertation, as a result of reassessing the notion of affective polarization, meets the desiderata DISANALOGY, EVIDENCE and INTERVENTION. First, it allows us to accommodate the difference between claiming that one is in a state of mind and being in a state of mind. Second, it allows us to accommodate cases in which a belief attribution serves for very different purposes. Thus, it allows us to differentiate between two types of situations that seem radically different, those in which someone does not believe what she says that she believes and those in which someone expresses certain affective attitudes, and consequently it allows us to qualify our polarization measurement tools much more. This is an important step for being able to measure the pernicious type of polarization as soon as possible.

This chapter is structured as follows. In section 6.1, we make a first attempt at presenting the general framework of our interpretation of some of Wittgenstein's insights concerning the mind. In section 6.2, we present a rough taxonomy of mental states that can be traced in Wittgenstein's writings, and introduce a specific sense of the term 'description' that can be also traced along Wittgenstein's philosophical production and that will enable us to argue that certain types of mental state ascriptions does not have a descriptive function. In section 6.3, we delve into Wittgenstein's anti-descriptive approach to psychological vocabulary. In particular, we present three arguments that can be found among Wittgenstein's reflections throughout his philosophy. In section 6.4, we present what it means to be in a state of mind, such as that of beliefs, from this view by drawing on the notion of rule-following: to believe that p is crucially to follow certain rules. In section 6.5, we hinge on a related question, the question of how the meaning of our expressions is determined, to highlight one difference that is crucial in this dissertation and that is present in Wittgenstein's philosophy under different ideas: the distinction between the descriptive and the evaluative. In section 6.6, we introduce the notion of contextual authority that follows from the view of the mind provided and that will enable us to satisfy the goals of this dissertation.

## 6.1. A Wittgensteinian nondescriptivist approach: A first attempt

In this section, we present the conception of the mental in which we rely on in this dissertation to satisfy the requirements previously discussed, essentially those that have to do with conceiving mental states in a way that allows us to know people's mind without directly asking them. In particular, we present here our interpretation of Wittgenstein's anti-descriptivism regarding mental vocabulary. According to our reading, some of Wittgenstein's remarks count as, or provide inspiration for, a type of expressivism that is very far from the simple expressivism usually attributed to him (see section 5.3). Some essential ideas of our interpretation are the following. First, there is no asymmetry between the first- and the third-person regarding the function accomplished by mental state ascriptions: both first and third person mental ascriptions play an expressive rather than descriptive role (see section 6.4). Second, there is no significant difference between belief-like and desire-like mental states regarding their link to action (see section 6.4). Third, there is a crucial difference between the rules one says one follows and those one actually follows.

Although we inevitably think that our interpretation of Wittgenstein's ideas is the right one, our purpose here is by no means exegetical. Rather, we will simply offer a set of arguments and observations that we find in Wittgenstein's writings and that seem useful to accomplish the goals of this dissertation.

According to our interpretation of a large group of Wittgenstein's insights, having a belief, a wish, an expectation, a hope, etc., is far from having something inside our heads or aiming exclusively at a set of observable behaviors that are the result of some inputs. Rather, it is having a set of conceptually articulated commitments that are linked to certain courses of action. Dain calls them *attitudes*, as opposed to being of the opinion that so-and-so (Dain 2019). For instance, believing that p is to assume the commitments conceptually tied to p, its grammar, its logical relations, as well as the courses of action linked to the conceptual relations of p. These conceptual relations depend on the logic of our language, on the rules that govern it, and in that sense they are objective: even if one is not aware of the conceptual connections of p, that one believes that p will depend on one having the commitments tied to p (among which it is self-attributing such a belief, but not only and not necessarily so). Crucially, the question of what the conceptual links to which one is committed are is highly context-dependent, and is determined by our practices. The grammar of 'being racist', for example, is determined by the practices in which this expression is used, as well as the courses of action that are usually carried out in such practices. Thus, if someone performs the courses of action that are normatively linked to being racist given our practices, then she is a racist, regardless of what she says about herself. In that sense, there is no gap between believing that p and behaving like one who believes that p.

Note, however, that the matter is not reduced to behavior in the sense of behaviorism: what is crucial here is *normativity*, the set of rules that links certain actions with certain expressions, in particular a conception of normativity which has in its core the possibility of error. Believing that someone is inferior due to the culture she belongs to does not depend exclusively on what one says about one's beliefs; it depends on following the rule according to which one believes such a thing, that is, it depends on behaving in a way that is in accordance with the rule of believing that someone is inferior because of the culture to which she belongs. Let's say it once again: these behaviors, verbal and nonverbal, internal and external, are normatively determined.

This interpretation draws primarily on the work of Villanueva, who has recently pointed to a new path in the debate between those who think that there is a rupture between Wittgenstein's early and late writings, i.e., the so-called Old Wittgensteinians, and those who think that there is a continuity between both writings mainly because the *Tractatus* offers simply a reductio of the representationalist theory of language, i.e., the so-called New Wittgensteinians (see Crary & Read 2000). Villanueva holds that there is a continuity between both earlier and later writings of Wittgenstein, but in virtue of some ideas offered in Wittgenstein's early production being developed and enriched in the mature stage of Wittgenstein's thought (Villanueva 2019). We follow this trail here, also suggested by the following Drury's quote:

> When Wittgenstein was living in Dublin and I was seeing him constantly he was at that time hard at work on the manuscripts of the Investigations. One day we discussed the development of his thought and he said to me (I can vouch for the accuracy of the words): 'My fundamental ideas came to me very early in life'. (Drury 1973: ix)

## 6.2. Taxonomy and "description"

Through a large part of his work, Wittgenstein claims that the psychological vocabulary exhibits a different logical behavior from that exhibited by superficially analogous expressions: although the expressions 'believing', 'knowing', 'understanding', 'thinking', 'expecting', 'intending', 'wishing', 'imagining', 'feeling', 'loving', are superficially similar to expressions like 'writing', 'eating', 'running', etc. –because they indicate something like a relationship between a subject and something else, their grammar, their function, their logical behavior, is radically different (Wittgenstein RPP I § 284, 472), among other things because the psychological vocabulary does not point to any activity (Wittgenstein RPP II § 193).[2]

In accordance with the spirit of his method, Wittgenstein did not offer any exhaustive taxonomy of mental states –although he was close of it in his mature stage. However, his grammatical investigations allow us to reconstruct something like a map that, although with limits that blurred and open to modifications, it

[2] Of course, there are important differences between the logic of those psychological expressions. For example, saying that someone believes that p is, among other things, to say that it is not true that p, or at least not take issue with the truth of p. On the other hand, saying that someone knows that p is endorsing p too.

distinguishes different language games that are usually grouped under the 'mental' label. On the one hand, Wittgenstein distinguishes mental states such as believing, thinking, knowing, understanding, wishing, hoping, etc., which he calls "dispositional" in part of his late production (essentially in PI § 149, RPP II §§ 43, 45, 281, LWPP II p. 9, p. 12), and sometimes also refers to as "hypothetical mental process", "hypothetical mechanism", "curious mental mechanism" and "special mental state". On the other hand, there are mental states such as sensations, emotions, moods and images, which he calls "states of consciousness" (PI § 149). Within the category of states of consciousness we can also distinguish between sensations, emotions and sensory impressions, that do not exhibit the same conceptual particularities.

We will be interested here just in the mental states that Wittgenstein calls dispositional, because they include those we are interested in for measuring polarization. In this regard, it is important to clarify from the outset that dispositional states of mind are not mere psychological tendencies, which is what is sometimes called 'dispositions' in the recent literature. Dispositions, in this sense, are psychological facts that do not allow for error. That is, they are 'automatic' inclinations to do certain things when something happens, such as the tendency of sugar to dissolve in water. The states of mind that Wittgenstein calls dispositional are normative and, therefore, are not mere psychological inclinations. Henceforth, with 'dispositional' we refer to Wittgenstein's sense.

Wittgenstein insists on the functional diversity of language, that is, the possibility that a word appears in two or more different language games. In this sense, someone might think that it is not accurate to say that, for Wittgenstein, mental state ascriptions are not descriptions (see, for example, Macarthur 2010). In fact, Wittgenstein sometimes uses the predicate 'being a description' for a mental state. How, then, can Wittgenstein be attributed a nondescriptivism view about the mental? The first thing to note is that Wittgenstein draws our attention to the confusions that natural language can lead us to, but that's compatible with our natural way of using language. In other words, the problem is not to say that a mental state is a description, but losing sight of what we mean with that claim. Second, a particular notion of describing can be traced in Wittgenstein's philosophy, which remains from his earliest writings to his late production (Villanueva 2019). This specific sense of description is all that is required to claim that, for Wittgenstein, dispositional mental states are not descriptions. And that's compatible with other senses of 'description' under which a dispositional mental state might be a description. What is the particular notion of describing according to which dispositional mental states are not descriptions? Let's introduce it.

In the *Tractatus*, Wittgenstein offers a picture of meaning, the pictorial theory of language, which is aimed at grasping the relationship between language and reality. According to this picture, propositions, which are linguistic entities, have meaning and can be true or false because, like a picture, represent or describe something, specifically states of affairs that can be the case.

A state of affairs is a particular combination of objects, which reaches the status of fact when it is the case. And each object of a state of affairs is individuated by its possibilities of combination with other objects, that is, by the way in which they can appear in other states of affairs, which is its form (Wittgenstein T § 2.011, § 2.0123 and § 2.0141). For instance, the state of affairs *the glass is on the table* is a fact, that is, it is a particular combination of objects that is the case. *Glass* and *table* are the objects they are because of their possibilities of combination with other objects, i.e., because of their form. Moreover, the states of affairs inherit their logical form or structure from the form of the objects composing them.

Propositions have the capacity of representing states of affairs because they are in an *internal relation* with the states of affairs they represent, where this means that each constituent of the proposition shares the possibilities of combination with the object of the state of affairs that it corresponds to. Thus, the proposition "the glass is on the table" describes the fact *the glass is on the table* because each ingredient of the proposition, e.g., 'glass', shares the possibilities of combination with the object they represent, e.g., *glass*. Thus, only *particular combinations of objects* that can be the case can be described, and propositions describe such states of affairs by virtue of the internal relation established between the constituents of the proposition and the constituents of the state of affairs that the proposition describes.[3]

[3] It is important to note that the theory of language offered in the *Tractatus* is not a standard

What is excluded by the law of causality cannot be described, says Wittgenstein: "What can be described can happen too: and what the law of causality is meant to exclude cannot even be described" (Wittgenstein T § 6.362). In the *Tractatus*, just bipolar propositions (those that can be true or false by virtue of the state of affairs they describe) belong to what can be *said*; everything else[4] belongs to the realm of what can be shown. Only what can be described can be said. If a sentence does not describe a state of affairs that can be the case (e.g., "It is morally right to help people"), then that sentence does not express a proposition, but a pseudoproposition –for Wittgenstein, the term 'pseudoproposition' is not pejorative at all: the most important realm of human reflection is formed by statements that are not propositions. Thus, in the *Tractatus* only particular combinations of objects not excluded by the law of causality, i.e., what can be said, can be descriptions.

In his later writings, Wittgenstein insistently returns to a notion of 'describing' which is very similar to that introduced in the *Tractatus*. According to this notion, descriptions express empirical propositions, they are representations of particular distributions of objects in space and time that are subject to causal restrictions. As Child puts it, "describing" is "a definite activity, which involves observation and the assessment of evidence" (Child 2017: 470). Villanueva presents the following three of Wittgenstein's passages to support this idea:

> So it depends wholly on our grammar what will be called possible and what not, i.e. what that grammar permits. But surely that is arbi-

descriptive theory of language, that is, this theory does not maintain that meaning is determined by a 1 to 1 relationship between a proposition and a state of affairs. Wittgenstein insists that objects are individuated by their possibilities of combination, and we can only distinguish two objects once we have all the combination possibilities given (Wittgenstein T § 2.0123, § 2.0124). In this sense, the meaning of a proposition can be known only when there is already a complete and closed system of possibilities of combination of the logical objects. Therefore, propositions do not acquire their meaning nor are they true when they describe a brute fact of the world; the meaning is determined by the system itself.

[4] We are aware of the distinction between nonsense and senseless, and of the possibility that there could be things that cannot be said nor shown. However, these possibilities are irrelevant to our purposes here.

> trary! Certainly; but the grammatical constructions we call empirical propositions (e.g. *ones which describe a visible distribution of objects in space and could be replaced by a representational drawing*) have a particular application, a particular use. And a construction may have a superficial resemblance to such an empirical proposition and play a somewhat similar role in a calculus without having an analogous application; and if it hasn't we won't be inclined to call it a proposition. (Wittgenstein PG § 82)

> If you trained someone to emit a particular sound at the sight of something red, another at the sight of something yellow, and so on for other colours, still he would not yet be describing objects by their colours. Though he might be a help to us in giving a description. *A description is a representation of a distribution in a space* (in that of time, for instance). (Wittgenstein PI, Part II ix)

> It positively seems to us as if pain had a body, as if it were a thing, a body with shape and colour. Why? Has it the shape of the part of the body that hurts? One would like, e.g., to say ”I could describe the pain, if only I had the requisite words and elementary concepts”. One feels: all that is lacking is the necessary nomenclature (James.) As if one could even paint the sensation, if only others would understand this language.–*And one really can give a spatial and temporal description of pain.* (Wittgenstein RPP I § 695)

We would like to strengthen the thesis that there is a notion of description –essentially linked to the distinction between saying and showing– which remains throughout Wittgenstein's writings, by adding other paragraphs from Wittgenstein's mature production:

> Couldn't different interpretations of a facial expression consist in my imagining each time a different kind of sequel? Certainly that's often how it is. I see a picture which *represents a smiling* face. What do

> I do if I take the smile now as a kind one, now as malicious? Don't I imagine it with *a spatial and temporal context* which I call kind or malicious? Thus I might supply the picture with the fancy that the *smiler was smiling down at a child at play*, or again *on the suffering of an enemy*. This is in no way altered by the fact that I can also take the at first sight gracious situation and interpret it differently by putting it into a wider context. If no special circumstances reverse my interpretation I shall conceive a particular smile as kind, call it a "kind" one, react correspondingly. *That is connected with the contrast between saying and meaning*. (Wittgenstein PG § 128)

> One will also be able to say: What this *description* says will get its expression somehow in the movement and the rest of the behaviour of the child, but also in the *spatial and temporal surrounding*. (Wittgenstein RPP I § 1067)

> Soulful expression in music. It is not to be described in terms of degrees of loudness and of tempo. Any more than is a soulful facial expression describable in terms of the *distribution of matter in space*. Indeed it is not even to be explained by means of a paradigm, since the same piece can be played with genuine expression in innumerable ways. (Wittgenstein CV: p. 93)

> It is possible to describe a painting by describing events; indeed that's the way it would be described in almost every instance. "He's standing there, lost in sorrow, she's wringing her hands..." Indeed, if you could not describe it this way you wouldn't understand it, even if you could describe the distribution of colour on its surface in minute detail. ((Picture of the man ascending the mountain.)) (Wittgenstein RPP II § 385).

What can be said, distributions of objects in space and time, empirical propositions, what belongs to the law of causality: all of these count as 'descriptive' under a particular notion of being a description. On the contrary, what belongs to the realm of morality, aesthetics, logic, meaning and, in particular, dispositional mental vocabulary, does not count as a description, under this sense of 'description'; but to what can just be shown.

## 6.3. Wittgenstein's approach to dispositional mental states: anti-descriptivist arguments

Take the following paragraphs from *Tractatus*:

> At first sight it appears as if there were also a different way in which one proposition could occur in another. Especially in certain propositional forms of psychology, like "A thinks, that p is the case", or "A thinks p", etc. Here it appears superficially as if the proposition p stood to the object A in a kind of relation. (And in modern epistemology (Russell, Moore, etc.) those propositions have been conceived in this way.) (Wittgenstein T § 5.541)
>
> But it is clear that "A believes that p", "A thinks p", "A says p", are of the form "'p' says p": and here we have no co-ordination of a fact and an object, but a co-ordination of facts by means of a co-ordination of their objects. (Wittgenstein T § 5.542)

In § 5.541 Wittgenstein refers to the relational theory of mental states, according to which sentences of the type "believing (expecting, hoping, desiring, etc.) that p" express a dual relation with an entity, i.e., the proposition p, or a multiple relation (Russell 1986). In § 5.542 he rejects any relational theory by saying that "A believes that p" is not a proposition, i.e., it does not represent any state of affairs.[5] On the contrary, "A believes that p" is a pseudoproposition like "'p' says p", that is, it is a sentence that points to an internal relation, and its truth-value does not depend on the empirical domain. The sentence "'p' says p" points to the

[5] Frank Ramsey also rejected the relational theory of mental states. Incidentally, he was the only one who, in Wittgenstein's words, really understood the *Tractatus*.

internal relation established between the proposition ‘p’ and the state of affairs p represented by ‘p’. This internal relation, which enables ‘p’ to describe p, is necessary, i.e., it belongs to the logic of language itself; it cannot be described (Cerezo 1998, 2003; Forero-Mora & Frápolli 2021). That kind of sentence does not express a proposition because it does not describe a state of affairs: it points to the logical rules that necessarily relate some propositions to others because of their logical structure. Hence, if “A believes that p” is the form of “‘p’ says p”, it follows that “A believes that p” does not describe anything that can be true or false; there is no fact that makes such a sentence true or false. In that sense, it points to the internal relations, to the logic of the system itself, and hence belongs to the realm of what cannot be said, like logic, ethics and aesthetics.

If belief ascriptions and self-ascriptions don’t describe state of affairs, we start to be in a good position to accommodate those cases, crucial for this dissertation, where there is possibility of error regarding mental self-ascriptions. In particular, those cases where someone sincerely self-ascribes a belief and she does not have such a belief, and where someone sincerely self-ascribe a belief and her claim expresses her practical attitudes, what can be expected from her. Both types of possibility of error[6] undermine descriptivist approaches to belief ascriptions, and the recognition of this is crucial to the question of how to measure ideological and affective polarization, and to understand the studies and tools we already have for measuring both types of polarization.

In the immediately following proposition, § 5.5421, Wittgenstein says “This shows that there is no such thing as the soul –the subject, etc.– as it is conceived in superficial psychology”. This statement might be better understood in relation to another remark of his mature stage:

> The soul is said to leave the body. Then, in order to exclude any similarity to the body, any sort of idea that some gaseous thing is

[6] Although we are calling ’error’ both types of situations, the first one is quite different from the second. The second type of situation shows that our everyday uses of the term belief are varied (and include the expression of practical commitments) and cannot be captured by the technical concept of belief. It is, however, a type of error insofar as one might think that through a self-ascription of a belief, the subject just gives information exclusively about the belief that she self-ascribes.

> meant, the soul is said to be incorporeal, non-spatial; but with the word ”leave” one has already said it all. Shew me how you use the word “spiritual” and I shall see whether the soul is non-corporeal and what you understand by “spirit”. (Wittgenstein Z § 127)

The superficial psychology, as Wittgenstein calls it, conceives the soul as a thing that, despite being incorporeal and non-spatial, *leaves* the body. ‘Leaves’, however, is a dyadic predicate that belongs to that can be described, as the state of affairs *Manuel leaves the room.* The proposition ‘Manuel leaves the room’ can be true or false because it describes a particular distribution of objects. But the mind, in particular dispositional mental verbs, point to internal relations, and internal relations cannot be described. Hence, superficial psychology misunderstands the logic of ‘soul’, ‘mind’, ‘believing’, etc.: it places it in the realm of what can be described, but it does not belong to that field (see Ryle 2009 for a similar thesis). The crucial insight behind the last remark is the distinction between what can be said and what cannot be said. Thus, in the *Tractatus* Wittgenstein already offered arguments against the descriptive nature of dispositional mental verbs, an idea that can be found in his later writings.

In *Philosophical Investigations*, Wittgenstein faces a view offered as a solution to the problem of intentionality, that is, the problem of how mental states are connected to their contents, which appear to be reached up by the world in a very peculiar way. This view, which is a kind of relational theory, sometimes referred to as *the harmony between thought and reality* (PI § 429; PG § 88, § 95, § 112, § 113; Z § 55), maintains that the truth of a belief ascription is supported by a fact that it reaches up to our mind. The general idea is that when someone believes that p, she is in something like an incomplete mental state, which is completed by the fact represented by the proposition p. For example, if I believe that the glass is on the table, the fact that the glass is on the table is what will make my belief true, as if the fact were a solid cylinder that fills a hollow cylinder into which it perfectly fits (PI § 439). The world reaches up to our minds.

According to Wittgenstein, this picture about dispositional mental states is ‘weird’, and wrong. How is it possible for me to know, before it happens, the fact that will make my belief true, or the fact that will satisfy my desire, or my expectation? (PI § 437). How can a fact be contained in my belief? What is this kind of relation between the mind and the world? What is the nature of the objects that mental states point to? (PI § 428). Wittgenstein's solution to these puzzles is to deny, as in the *Tractatus*, that dispositional mental states, like beliefs, are relational. Belief ascriptions do not describe a relation between a subject and an intentional object, but point to the grammatical connection between two propositions. The relation is conceptual, says Wittgenstein (PI § 445; PG § 88; Z § 55). The connection between 'I believe that the glass is on the table' and 'The glass is on the table' is not a factual relation between two things. Instead, the claim "the fact that the glass is on the table is what makes the belief that the glass is on the table true" is just a grammatical insight about the concept of 'believing'; a reminder of how it works. Dispositional mental states point to the conceptual relations between propositions, i.e., to the commitments and conceptual links that logically exist between them. If I say that I believe that p, I cannot say that my belief is false after asserting that p is the case, because the logic of 'believing' demands such a commitment from me. And a similar reason is behind Wittgenstein's claim that the utterance of the sentence "I know I am in pain" express a nonsense: doubt is logically excluded from the first-person perspective in mental state self-ascriptions (PI § 246).[7] The attitudes we have been talking about in the first chapters, especially in chapter 4 where we have reviewed the evidence on how we get polarized, are these dispositional states of mind. When we want to get an idea of the distribution of public opinion with respect to a particular issue, we are exploring the distribution of these dispositional states of mind, of these attitudes, in a group.

Acero and Villanueva have argued that not only the diagnosis of the gram-

[7] This remark is taken as an example of Wittgenstein's rejection of the epistemic explanation of first-person authority (see Hacker 2005). We agree on that. However, it is important to note that stressing that doubt is logically excluded from the first-person perspective is compatible with the idea that the speaker's sincerity is not enough to guarantee the truth of her mental self-ascription: it is a nonsense that a speaker expresses doubt about whether she is in the mental state she says to be in, but despite the fact that she cannot logically doubt it, she can fail in identifying the mental state she is in.

matical connection points out that dispositional mental ascriptions do not have a descriptive function, but also that it is essential to the diagnosis itself to conceive these mental ascriptions as *expressions* of the corresponding thoughts of the agents, together with other conditions: semantic innocence, language as the vehicle of the thought, and systematicity (Acero & Villanueva 2012: 137-138). So, it can be argued that this diagnosis is nothing else but a complement to the diagnosis offered in the *Tractatus* against the relational conception of mental states.

Finally, Wittgenstein pointed out that belief ascriptions, unlike what is describable, do not have a genuine duration (RPP II §§ 51, 178; Z §§ 46, 82; see also Ramos 2001; Villanueva 2019). Our beliefs do not cease when we concentrate our attention on a task (RPP II § 45), nor are temporarily measurable. To put the point somewhat differently, the time during which we had a belief or the particular moment in which we began to have it cannot be exhaustively delimited temporarily. In this sense, Wittgenstein says that beliefs, and other dispositional mental states, do not have a genuine duration: if we believe, for instance, that your social identity shapes the type of actions that you *perceive* as possible, we have that belief latently; it is not possible to measure the exact time at which we believed it or to point the specific minute in which we began to believe it. Of course, we can describe the event that *made us* believe that, under a particular way of talking. But the duration of the belief itself is very different from the temporality of a physical event. In that sense, they do not belong to the realm of which can be described, that is, distributions of objects in space and *time*.

As we have said, belief self-ascriptions are usually associated with ideological polarization: it is commonly assumed that respondents self-report their beliefs through their responses to the surveys designed to measure ideological polarization. However, as we have been pointing out, it is not unusual for someone to sincerely say that she believes something and it is the case that she does believe such a thing, and other times someone says that she believes something and she is just expressing certain affective attitudes rather than reporting that particular belief. Both situations call into question that people have authority regarding their mental self-ascriptions.

We will argue that all we have is contextual authority. And this is crucial both for measuring ideological polarization and affective polarization. If we want to know the contents believed by a population, a more precise way of knowing this is to measure polarization indirectly, trying to measure not the beliefs that the population self-ascribes but those that they express to have. And if we want to measure the type of polarization that has to do with the level of radicalism, a more precise way to do it is to measure polarization indirectly, trying to measure the affective attitudes, those with an especial link to action, that the population expresses to have and that are connected to a certain level of radicalism.

## 6.4. The picture of the mind: Following rules

Again, when someone sincerely self-reports a belief or a feeling as an answer to a question of a tool used to measure ideological or affective polarization, she can be in error in two different ways. On the one hand, she can fail to identify her actual mental state: even though she sincerely self-ascribes it, she can be wrong and not being in that mental state. On the other hand, she can express her practical commitments through her belief or feeling self-ascription rather than simply report it. How can we know whether one of these two possible situations is the case?

As we have seen, the relation between dispositional mental verbs and the propositions to which they are directed at is an internal one, i.e., a logical, grammatical or conceptual connection. The correctness of a mental self-ascription or a mental state attributed to someone, i.e., its truth-value, depends on whether the subject of the mental ascription is committed to what logically follows from having that particular mental state. Take the following remark as an example of that:

> The sentence ”I want some wine to drink” has roughly the same sense as ”Wine over here!”. *No one will call that a description*; but I can gather from it that the one who says it is keen to drink wine, that at any moment he may take action if his wish is refused –and this will be called a conclusion as to his *state of mind*. (Wittgenstein RPP I §

469, our emphasis)

If I sincerely claim "I want some wine to drink", then I'm committed to what conceptually follows from it, that is, for example, that I'm keen to drink wine, and that I take action if my wish is refused. 'What conceptually follows' means what is logically allowed and therefore does not constitute a violation of the rules of the language game in which the expression is embedded (see Dain 2019). If, after claiming "I want some wine to drink", it turns out that I'm not willing to drink wine, or I get angry because they bring me a glass of wine, or I am happy if they tell me that there is no wine (without any other relevant information), then it will be false that I wanted some wine to drink, because the links between my mental self-ascription and other statements and courses of action connected to it (its logic, its meaning) have failed. And in the other direction it works the same way: if it turns out that I am willing to drink wine, I get angry if they tell me there is no wine, I am very happy when they bring me a glass of wine, etc., then I want some wine to drink, no matter what I explicitly say about my wishes. This is one of the points of Wittgenstein's analysis of situations in which, for example, I expect him to come: "What's it like for me to expect him to come? I walk up and down the room, look at the clock now and then, and so on" (PI § 444). These actions are logically permitted in such circumstances. Dispositional mental states are linguistically articulated and linked to certain courses of action.

This claim is not limited to first person mental ascriptions. 'Believing', in the first and the third person, is a dispositional mental state. How can we know other people's mental states? "This is shown me in the case of someone else by his behavior, by his words. And specifically by his expression 'I believe' as well as the simple assertion" (PI II, x, pp. 191-192; LWPP II p. 12). However, as noted, this should not be understood as a kind of behaviorism: 'thinking', 'believing', 'understanding', etc., is not just behavior (RPP II 12), "but a state of which this behavior is a sign" (PG 41). Certain courses of action are linked to each mental state, i.e., to each set of conceptual commitments; they form part of its logic, its meaning. And, in that sense, there is no asymmetry between the first and the third person: "What about my own case? Do I study my disposition in order to make the assertion or the utterance "believe"? –But couldn't I make a judgment about this disposition just like someone else? In that case I would have to pay attention to myself, listen to my words, etc., just as someone else would have to do" (PI II, x, pp. 191-192; LWPP II p. 12).[8] Presumably, this is what supports Manne's discussion of incels. Sincerely claiming that one believes that women are mindless creatures is not necessarily enough for that self-ascription to be true. The rule one says one follows might not be the rule one actually follows, as in the case of incels. Moreover, one can self-ascribe a belief or a feeling and, through such a self-ascription, one can express that one follows a very different rule, one especially linked to action, as in the Obama's case presented at the beginning of this chapter. It could even be argued that, at least in some cases where incels self-attribute the belief that women are mindless beings, in fact what they are doing is expressing their affective attitudes, i.e., what can be reasonably expected from them, rather than self-ascribing a belief.

As we will see in section 7.1 of the next chapter, this second case is one of those cases where we express certain attitudes with an especial link to certain courses of action. Cases where we use language in an evaluative way.

One of the most radical and beautiful conclusions that follows from these observations about mental state ascriptions is that the mind is nothing, in an ontological sense. But it is still something: normativity. Mental verbs are expressions that we use to talk about what someone is expected to do, that is, her commitments, conceptually articulated with certain courses of action by virtue of the normativity of our language and practices. Psychological verbs are expressions that we use in certain practices, in certain language games, and its correct application depends on whether they are in accordance with the rules that govern

[8] One of the famous paragraphs used to attribute to Wittgenstein an asymmetry between the first and the third person is RPP § 63, where Wittgenstein explicitly says that psychological verbs in the third person of the present are identified by observation, while not so in the first-person. However, Wittgenstein immediately notes that this is "((not quite right))". Of course, there are important conceptual differences between the first- and the third-person, but the distinction between expressing and describing cannot be one of them, because it will leave unexplained the remarks previously discussed. Dispositional mental verbs, both in the first- and the third-person, are expressions rather than descriptions.

that practice or game. Nothing else. There is no nature to discover, there are no mental processes to investigate in the strong sense of 'mental process'. All we can do is to study the logic of psychological vocabulary, the internal relations to which psychological concepts point to, to try to undo the associated ontological misunderstandings and unravel their semantic role. And the matter is not, as it is sometimes said, that this analysis is simply about the semantic function of psychological verbs and not about the ontological issue. Nothing could be further from the truth. Wittgenstein's conceptual analysis aims precisely at showing the confusion that superficial grammar has led us to: research on the nature of mind is flawed from the beginning; all there is it is language and following rules, i.e., normativity. The mind is nothing, neither internal nor external to the subjects. And yet, mental vocabulary is ineliminable.

There are semantic positions about certain uses of language, the evaluative uses of language, that start from a rejection of the postulation of strange ontological entities, and that explains the use of evaluative language in terms of the commitments expressed. These positions are very close to our just introduced interpretation of a number of Wittgenstein's remarks on dispositional mental states, and could provide more tools in order to distinguish between the two types of error regarding mental state ascriptions. One of these positions, as we shall see in the next chapter, is expressivism, a semantic theory about the meaning we express through the evaluative use of language, that will enable us to explain the type of commitments we are associating with affective polarization. As we will see, through the evaluative use of language we express our attitudes especially linked to action, that is, what can be reasonable expected from us. In this sense, through the evaluative use of language, we will argue, we express our affective attitudes connected with our level of credibility in certain beliefs of the group we identify with. These attitudes are the relevant ones to measure affective polarization, as we reassessed the concept. In the next section we will make a first approach to the distinction between the descriptive and the evaluative from Wittgenstein's philosophy.

## 6.5. A Wittgensteinian approach to meaning

A fundamental piece to see more clearly this picture of the mind is the question of how the meaning of our expressions is determined. In particular, the idea that there are no extra linguistic facts beyond our human practices supporting the meaning and the truth of our claims is crucial, as it is that we can use language in different ways and with different purposes: sometimes we use language to *describe* our surroundings, for which the notion of fact is especially helpful in a sense to be qualified in section 6.7, and sometimes we use it to *express*[9] our commitments, our attitudes, our mental states. We think that both ideas are present in Wittgenstein's philosophy from the beginning of his production.

In the *Tractatus*, the meaning of a proposition is determined by the state of affairs it represents, but propositions represent states of affairs due to the internal relation established between both things, and these internal relations belong to the system as a whole. That is, it requires that all objects, each one individuated by its possibilities of combination with other objects, are given in advance. In other words: the meaning of each proposition is given by the whole system, by its logic, which is itself not describable. Moreover, in the *Tractatus* we have the distinction between external vs. internal relations, or the distinction between what can be said and what can be shown. These things correspond to different functions of language. What can be said is what is describable, i.e., what points to distributions of objects, while what can be shown is that which cannot be described, i.e., what points to the logic of the system, its rules, for example claims about ethics, aesthetics, logic, meaning, psychology, etc.

In the mature stage of Wittgenstein's thought, the meaning of a proposition

[9] Note that this notion of 'expression' is different from what it is normally attributed to Wittgenstein (see, for instance, Barroso 2015: 52-58), i.e., the linguistic manifestation of a subjective experience (e.g., pain) that replaces natural expressions like cries and moans and therefore has the same semantic function, i.e., it is not truth-apt. That's the notion in which Bar-On bases her attribution to Wittgenstein of simple expressivism (see also Wright 1998). 'Believing', however, does not replace any primitive expression; its expressive function consists in acquiring and showing the conceptual commitments and the linked courses of action that can be expected from someone who believes that p, because 'she believes that p' does not describe anything, but points to internal relations.

is determined by the practices it belongs to, i.e., its use in a specific language-game, which can be illustrated with the following quote from one of Wittgenstein's writings prior to what is considered his mature period: "if we had to name anything which is the life of the sign, we should have to say that it was its use" (Wittgenstein BB § 4; see also PI § 43). Thus, meaning is not determined by the things we refer to; it has to do with a form of life, which is the general context in which language-games are inserted (in *On certainty*, Wittgenstein often refers to language-games as a structure, system, set of rules to act). "To imagine a language means to imagine a form of life" (PI § 19). Moreover, Wittgenstein distinguishes between empirical vs. grammatical propositions among other different uses of language. Empirical propositions, as we have seen, are those aimed at describing a distribution of objects in space and time, i.e., states of affairs that can turn out to be facts, while grammatical propositions are roughly similar to what he called pseudopropositions in the *Tractatus*. Thus, "The glass is on the table" and "I want some wine to drink" are two expressions that, in the suitable contexts, belong to radically different language-games: the former belongs to a descriptive one, while the latter belongs to a non-descriptive game, due to the rules of our language. Language-games are rule-governed areas of language, constitutive of human activity and, as noted above, are parts of a form of life. Let's say something more about the concepts of rule-following and form of life.

The problem that is generally alluded to by the question of rule-following in Wittgenstein's philosophy is expressed in § 201 (PI): "This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here". However, as Kripke points out, the problem posed by rule-following, as well as the solution offered to it, is primarily in §§ 138-242 PI (Kripke 1982). What is it that a certain action or a certain expression depends on to be in accordance with a given rule? This is fundamentally the problem of how we determine meaning: how does a certain word have its meaning?

Think of a case where someone is learning to follow a numerical series. In that case, one might make mistakes in different ways. For example, one could make random errors, in which no pattern is observed. But one could also make a systematic error (PI § 143). And it could also happen that one succeeds in following the numerical series to some extent simply by chance. In those cases, we would say that this person does not know yet how to follow the numerical series. We would only say of this person that she knows how to follow the numerical series when she is able to understand the circumstances in which one number should be written and not another, that is to say, when she is able to behave in a certain way. The same happens with concepts. To understand a concept is to understand the circumstances in which that concept has application, or the occasions in which someone misuses it. When someone uses a concept, she may not use it in accordance with the rule, or she may not have understood the rule, and so on. We need criteria to be able to discriminate between these different situations. Is it enough that someone sincerely claims that she knows how to follow the rule? "Suppose B says he knows how to go on -but when he wants to go on he hesitates and can't do it: are we to say that he was wrong when he said he could go on, or rather that he was able to go on then, only now is not?" (PI § 181). The answer is no. If someone sincerely claims that she knows how to follow a rule but it turns out that she does not know how to proceed, or she proceed erroneously, then she does not know how to follow the rule. The steps that are to be taken are determined by certain rules, that is, by the circumstances of application of, for example, a concept. And these rules depend on our practices.

Wittgenstein rejects the idea that there are facts in the world that determine whether a certain action is in accordance with a rule, i.e., he holds that meaning is not a matter of facts; all there is it is a form of life made up of practices governed by rules that allow sanction and correction. "There is a way of grasping a rule which is not an interpretation, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases." (PI § 201), ""So you are saying that human agreement decides what is true and what is false?"—lt is what human beings say that is true and false; and they agree in the language they use. That is not agreement in opinions but in form of life" (PI § 241). And that's the reason why we can think we follow a rule, for example that we believe that p, but in fact we don't follow the rule we think we follow: following a rule is a public matter, subject to correction. As Wittgenstein puts it: "And hence also 'obeying a rule' is a practice. And to think one is obeying a rule is not to obey a rule" (PI § 202). *In your acting, you express the rule you follow.*

In *On certainty*, Wittgenstein offers a similar picture, which can be briefly and roughly presented as follows. Some propositions, i.e., hinge propositions, constitute our structure, our system, our rules, our picture of the world from which we make judgments and from which it makes sense for us, and even it is right, to do certain things in certain circumstances (OC § 52). These propositions are certainties in the sense that they are assumptions we decided not to doubt about; they are the conditions of possibility for our other moves, the normative structure. These propositions are taught to us by acting, by practice, by instruction (OC § 95), and their justification comes to an end: we don't always have good reasons to believe them. The end of justification is an ungrounded way of acting, the set of practices that, together with the assumptions, constitute a form of life.

However, it is crucial to note that this picture comes in degrees. In other words, within the same form of life there are assumptions that we have decided not to question, along with other propositions and practices more or less established. But this situation is constantly changing (OC § 321). Wittgenstein uses the *riverbed* metaphor to exemplify this: In the same riverbed, some hard rocks can become sand and move downstream, and new rocks can also be generated.

Moreover, it is also important for our purposes to note that not all people belonging to a form of life (a set of practices and assumptions that give words their meaning) are engaged in all the practices that make up that form of life. For instance, the practices in which people use the expression 'Abortion is not morally wrong' might be different from the practices in which the expression 'Christian values must be followed', or the expression 'Abortion is legal in this country', appear. As it can be seen, the logic of some expressions makes them compatible with some and incompatible with others. For example, the expression 'Abortion is legal in this country' is compatible with 'Abortion is not morally wrong' or 'Christian values must be followed'. However, 'Abortion is not morally wrong' is not compatible with 'Christian values must be followed', because of the rules of the form of life they belong to. Thus, despite the fact that the meaning of the three expressions depends on the same form of life, two people belonging to the same form of life can be committed with different conceptual relations and courses of action, i.e., having different dispositional mental states: "the grammar of "believe" just does hang together with the grammar of the proposition believed" (OC § 313). For example, one may be committed with the conceptual relations, and its linked courses of action, of the proposition 'Abortion is not morally wrong' (e.g., she believes that abortion is not morally wrong), and another person may be committed with the conceptual relations, and its linked courses of action, of the proposition 'Christian values must be followed' (e.g., she believes that Christian values must be followed). The mind of someone is tied to the rules she follows, no matter whether she is aware of it or not. "When we first begin to believe anything, what we believe is not a single proposition, it is a whole system of propositions" (OC § 141). Saying "I believe that Christian values must be followed" not only makes possible that the speaker is not really in that mental state; it also gives information about what can be expected from the speaker, given the practices, the way of living, in which such claim are commonly stated.

There may be more or less shared practices between two people, and the practices can be of a different nature in virtue of how established they are. When two people share a sufficient set of practices, i.e., a background or standard, the information that one communicates to the other when talking about those practices is not especially linked to action. However, sometimes we rely on different and less hard rocks, i.e., different practices. For instance, if two people share the same scientific framework, or the same religion, then claiming that water boils at 100 degrees Celsius, or that Christian values must be followed, might give very little information to her interlocutor regarding what she can expect from the speaker. At least much less than if they don't share their standards. Besides, the practices supporting the truth of each claim are differently established in our form of life, and that also affects the type of information we express when making certain claims. That's part of what justifies the distinction between descriptive and evaluative uses of language within a system in which meaning is not determined by extra linguistic facts, as we will see in the next chapter.

So, as we have seen, when we sincerely claim that we are following a rule (e.g., when we sincerely claim that we believe that p) it could happen that we are not actually following such a rule. But we could also express not only that we are not following that rule, but that we are actually following a very different rule. These two possible situations correspond to our discussion concerning the two cases introduced at the beginning of the chapter. Thus, when the tools employed to measure polarization are used to find out what a population believes, we can encounter at least two possible situations. One is that someone sincerely says that she believes that p but does not believe it. And the other is that someone sincerely says that she believes that p and thereby expresses her commitments especially linked to action, her affective attitudes. This second possibility, opened by the dispositional view we have introduced in this chapter, is relevant to measure the kind of attitudes linked to affective polarization. Through their answers to some of the questions used to measure affective polarization, as we will argue, participants express certain attitudes with a special link to action, closely tied to their level of credibility in the core beliefs of the ideological group they identify with. To reach this conclusion, we need to complement the approach offered in this chapter with a particular approach to what we do when we use the language in an evaluative way (chapter 7).

## 6.6. Contextual authority

In this section, we introduce the notion of *contextual authority*, according to which there are contexts in which there is a presumption of authority regarding a mental self-ascription, contexts in which the speaker exhibits a strong authority, and contexts in which there is neither strong nor presumptive authority (Villanueva 2014). In the last case, the speaker might not be in the mental state she says to be in, and might express her affective attitudes, her way of living. The idea, supported by the Wittgensteinian picture introduced through the previous sections, is that what determines the truth of a mental self-ascription is its logical compatibility with certain propositions and courses of action. In particular, the truth of a mental self-ascription depends on its compatibility with the contextually salient features of the situation in which the self-ascription is made in virtue of the rules governing our language. Besides, the meaning expressed through a mental self-ascription is highly context-dependent as well, and in some cases a mental self-ascription serves to express one's practical commitments. Sincerely uttering "I believe that p" or even just "p", in the suitable context, is a criterion for ascribing the speaker the belief that p. However, it is neither the only criterion, nor necessarily the most relevant; it depends on the particular case. To put it another way, mental state self-ascriptions are not true or false considered in the abstract, but their truth-value, as well as the meaning expressed through them, has to be contextually determined. This view is compatible with a recent approach to first-person authority that takes it as an interpersonal, social norm, that is, the norm we follow when we defer to people's communicative expressions about what they feel, think, etc. (Borgoni forthcoming; Frápolli 2019; Navarro-Laespada & Frápolli 2018). In many contexts, we don't follow this social norm, and have good reasons to do it.

Consider again the cases CASE 1 and CASE 2 introduced in the previous chapter. In CASE 1, if we take Oscar's self-ascription "I believe there should be unrestricted freedom of expression" as true, then some salient features of the case remain unexplained. For instance, it makes no sense that Oscar frequently insults people who make jokes on Twitter and threatens to report them for hate crimes, or that he answered in a significant number of vignettes saying that someone should not say what she said. Believing the proposition "there should be unrestricted freedom of expression" binds us to a set of commitments and courses of action which are incompatible with, for example, answering a survey selecting the option someone should not say what she says. On the other hand, if we take Oscar's self-ascription as false, then most of the salient features of the case make sense. Oscar's behavior and reaction in some situations show that he does not believe what he says he believes, because believing so is to have certain conceptual commitments linked to certain courses of action.

One may object that the sincere self-ascription may itself be also a contextually salient feature of the context, and Oscar is sincere by making the self-ascription. Certainly, a sincere self-ascription is sometimes a relevant feature in assessing its truth. However, in this case it does not seem to be very relevant, especially because there is no particular reason to think that Oscar is right in his mental self-ascription despite being sincere. As we have seen, the rule one thinks one follows is not necessarily the rule one actually follows. Being in a particular state of mind is based on not violating the logic of being in that state. That explains our intuition in this case. Moreover, we are now in a position to see that, in this case, Oscar's belief self-ascription can be understood as a way of expressing his practical attitudes, his level of attachment to an ideology that has, as a core belief at a particular time, the idea that one has the right to say everything one wants whenever one wants to say it. Under this second interpretation, Oscar wouldn't be simply showing that he doesn't really believe what he says he believes, but also that he has a high level of credence in certain beliefs, in virtue of the other practical beliefs and attitudes that he expresses.

Consider now CASE 2. The explanation goes along the same line. If we take Kautar's self-ascription "I believe my friend's father is a racist" as true, then the most salient features of the case make sense, and that's a reason to take her self-ascription as true. For instance, it makes sense to think that Kautar, being an Arab descendant, has been quite exposed to situations in which others have behaved in a racist manner toward her, that is, she is more trained to detect racist situations. And none of the other contextually salient features of the case appears to be clearly contradictory with assuming that it is true that Kautar is in the state of mind she says to be in. Certainly, some features of the case, such as Kautar routinely opposing racism and distancing herself from people she considers racist, push a bit in the opposite direction. However, these features are not completely incompatible with Kautar believing what she says she believes, neither they are excessively salient features of the case. Her future behavior may be completely compatible with believing that her friend's father is racist. In that sense, the contextually salient features of the case are not incompatible with the truth of the doxastic self-ascription. Moreover, it is important to note that the contextually salient features of a case are very varied in nature. The relevant and salient features in considering whether someone has a particular mental state can be the claims that the speaker makes, the actions she performs, the thoughts, desires, expectations, psychological inclinations and sensations she has, her socio-normative position, the content of the mental state in question, the place in which the case develops, the general norms governing the practices of a society in a particular time, etc. (see chapter 7). The wide variety of relevant aspects in considering the truth of a mental self-attribution is what explains our intuition in this second case. Sometimes, the sincere mental self-ascription is the most contextually salient feature to determine its truth-value.

However, as previously noted, it is conceivable that two people A and B disagree on whether a third person C believes that p, in different senses. A and B might agree on what determines whether it is true that C is in a certain state of mind and then the disagreement will be based simply on the fact that one of them had missed some relevant features of the situation. In that case, they can easily settle the disagreement. However, it might also happen that, given A's way of living, i.e., the practices in which she is usually involved, she could give more weight to C's recent claims when deciding whether C is in a certain mental state. And it could happen that, given B's way of living, she could give more weight to certain C's facial expressions and actions when deciding whether C is in a certain state of mind. That is to say, it could happen that A and B disagree on what makes true that C is in a certain mental state, or even the meaning expressed through her claim. In this second scenario, A and B have different standards from which they attribute a mental state.

This possibility is precisely what enables us to argue that we can in fact adopt the political stance discussed in the last chapter of granting credibility to a person that self-ascribes a mental state, no matter how high the stakes become. Recall the tension: on the one hand, there are some arguments and sound evidence showing that we usually fail in identifying our own commitments and that we don't always have authority. On the other hand, in many contexts people exhibit authority, and in other contexts we have the obligation to believe what others say, we have to trust them, for political reasons but also for psychological ones: many studies from psychology show the benefits of trusting others (Yamagishi 2001; Yamagishi et al. 2002). The political stance discussed in the previous chapter, then, is exactly such a thing: a political stance about what we want to do, how we want to live, how we want our practices to be, etc. Given the social inequality and the continuous discrediting that people from different disenfranchised identity groups are exposed to, the socio-normative position of the person who self-attributes a mental state may be the most salient contextual feature in a particular case when determining whether that person is in the mental state she says to be in. And this type of evaluation is one of the things we actually do. There's nothing else. It is not necessary to postulate moral facts beyond our human practices, as Srinivasan does. Our motivation for a fairer society promotes practices in which a person's socio-normative position is sometimes the most salient contextual feature, and even the only relevant one, to consider as true her mental self-attribution. In a similar vein, it can be argued that more extreme cases of mental attribution, such as Juliet's case (section 5.5), simply press us to make an evaluation, i.e., to express the practices we are engaged with, our standards. In other words, these extreme cases reveal that some of our practices are not hard rocks; they are not much established and shared. All this is the outcome of acknowledging that dispositional mental state attributions are normative.

Thus, with the notion of contextual authority and the flexibility provided by the idea that mental ascriptions are made from a particular standard or way of living, both ideas supported by the Wittgensteinian picture of the mental offered in the two previous sections, we can accommodate our intuitions in CASE 1 and CASE 2, the political stance, and say something else about more extreme cases. Finally, it is important to stress that, given its context-dependence, the mental state expressed through one's verbal and nonverbal behavior will vary from a society to another, and even from a particular time to another and from a particular context to another in the same society. This lesson is central to our reassessed concept of affective polarization.

## 6.7. Conclusion

In this chapter, we have introduced a pragmatist and antidescriptivist view of some mental state ascriptions. According to this view, dispositional mental vocabulary is not in the business of describing distributions of objects in space and time. Rather, it has a normative flavor: its correct use is bound to certain rules. The normative character of dispositional mental state ascriptions implies that one can fail to identify the mental state in which one actually is, the rules one actually follows. The commitments or mental states one says one has can be different from those one actually has, and determining it is a highly context-dependent task. But not only that. By self-ascribing a dispositional mental state one might express certain attitudes especially linked to certain courses of action, beyond the courses of action linked to having or not having the belief one claims to have.

This picture can accommodate our intuitions in CASE 1 and CASE 2 without abandoning the political stance with the aid of the notion of contextual authority. Moreover, from the Wittgensteinian picture of the mind provided it follows that there might be a distinction between the mental state we self-ascribe, i.e., the rule we think we follow, and the mental state in which we actually are, i.e., the rule we actually follow. The rule we actually follow is expressed in our acting, that is, we show the commitments we acquire through our verbal and nonverbal behavior. Thus, we can track the mental state in which a person is in by focusing on the commitments she expresses through her use of language and the actions she performs, and therefore this picture satisfies DISANALOGY. But not only that. This Wittgensteinian approach can explain situations of the two types of possible errors when someone self-ascribes a mental state. On the one hand, this contributes to meet the desideratum EVIDENCE insofar as it can account for the evidence in this line. But, more crucially, this approach enables us to say something more precise regarding what someone might do through her mental self-ascriptions, and therefore allows to design tools that can measure affective polarization more accurately. If, as many studies show, it is true that the greater the level of polarization, the lesser the possibilities to depolarize, then this move also contributes to meet the desideratum INTERVENTION. Hence, this approach is in a better position than descriptivist ones to achieve the goals of this dissertation.

We would like to say something else before ending this chapter in order to connect it with the objective of the following chapter, which is to show that through the evaluative use of language we express our practical attitudes, the rules we follow and that tie us to certain courses of action. Note that from the Wittgensteinian picture of the mental that we have presented it follows that there are uses of language more closely linked to certain courses of action than others, given the practices in which they appear and given the rules governing those practices. For instance, there are situations in which our goal is to describe our surroundings. In those situations, we make claims such as "The glass is on the table", "The window is closed", "There is a person wearing a black shirt", "The water is boiling", etc. Within this practice, the main information conveyed by the utterance of those sentences is about the world. That is, we talk about distributions of objects in space and time, state of affairs that can be the case. In that sense, this language-game is a descriptive one. Of course, the expressions 'being about the world', 'being a distribution of objects', 'being a state of affairs' and others are simply predicates that can be attributed to other expressions that belong to this realm, to this language-game, like "The glass is on the table", in order to make a particular move. And the same goes for predicates like 'being a fact'. Thus, saying "it is a fact that the glass is on the table" is making a particular move within this language-game, conceptually linked with expressions like "It is true that the glass is on the table", "The glass is on the table is a state of affairs that is the case", and "The glass is on the table is a distribution of objects". These expressions, under this particular use, belong to widely established and shared practices in our form of life. In these situations, it can be the case that the speaker does not have the belief that she expresses to have through her claims.

However, since the meaning and function of words emerge from the practices in which they are used, this language-game works differently from the way other language-games do. For instance, claims such as "I hate when the glass is on the table", "Abortion should be legal", "The left is naive", "I would be angry if my daughter gets married with someone from Morocco", "People from southern countries steal our work", etc., in the suitable context, do not communicate information about how the world is, or not only that. Rather, they belong to an evaluative language-game, which is governed by other rules and involves different practices. In particular, the evaluative use of language is characterized by expressing a particular picture of the world especially linked to certain courses of action. In these situations, it can also be the case that the speaker does not have the belief that she expresses to have through her claims. But not only that. Besides the expression of that particular belief, the speaker expresses something else, her practical attitudes. Within this language-game, part of the information conveyed by the utterance of those sentences is not about the world, but about the expression of the speaker's practical commitments. For example, given the practices governing the use of the expression "People from southern countries come to steal our work", uttering such a sentence, in the suitable context, expresses information about what can be expected from the speaker. For example, it can be expected that someone who utters such a sentence in the suitable context will not try to help people from southern countries, will not mind that they are persecuted and beaten, will make other racist comments such as that people from southern countries receive a lot of financial aid, and so on. These attitudes expressed by the speaker, in contrast to the expressed belief that people from southern countries come to steal our work, cannot be discovered not to be the case. If someone expresses certain affective attitudes, then it cannot be the case that the speaker does not have those attitudes. While one can express a belief trough an assertion and be the case that the speaker does not have that belief.

As it can be seen, the evaluative use of language is much more linked with certain courses of action than the descriptive one. Of course, in the evaluative language-game we can also say something like "it is a *fact* that abortion should be legal". However, the logic of 'being a fact' here is not exactly the same that in the descriptive language-game: for example, it is not conceptually linked to there being a state of affairs, or a particular distribution of objects. Nevertheless, they do share that in both language-games saying that something is a fact is a way of saying that one acquires a high degree of commitment with the truth of the claim. And, of course, evaluative claims can be true or false; 'being true' simply has not exactly the same conceptual relations that it has in the descriptive game, because it doesn't include for example certain conceptual relations, such as being a state of affairs (see 7.3.1 for a complement of this discussion).

Take all of this just as a first sketchy introduction of the topic of the following chapter, which can be read as a unity together with this one. But keep in mind the following idea: when using language in a descriptive way, we convey information about how the world is and acquire a particular commitment to not explicitly denying that we believe that proposition. And from an external perspective, one can say that the proposition is true but it is false that the speaker believes that proposition. On the other hand, when we use language in an evaluative way, we convey information that expresses our own perspective, which is beyond our commitment to not explicitly denying that we believe that proposition. And, in this case, it is conceptually impossible that someone does not have the commitments she expresses to have. The remarks we have done along this chapter regarding attitudes are not an attempt to prove that the study of attitudes carried out so far from cognitive sciences is wrong. Rather, we have carried out a conceptual analysis of some of their assumptions that particularly affect the way in which affective polarization is measured.

# Chapter 7

# Attitudes and Evaluative Meaning

In a recent empirical study (Porter et al. 2019), already mentioned in chapter 4, researchers have tested how the correction of a misstatement affects people's political beliefs and preferences. In particular, they have tested the correction of the two following misstatements made by Donald Trump at some point:

> [Climate change] wasn't working out too well, because it was getting too cold all over the place. The ice caps were going to melt, they were going to be gone by now, but now they're setting records . . . they're at a record level. (Porter et al. 2019: 3)

> The [Paris Climate] accord would prohibit America from building new coal plants while giving permission to China and India to build them. (Porter et al. 2019: 3)

In both experiments, participants were randomly assigned to be exposed to a misstatement, or a misstatement and a correction of it, or neither. After that, they were asked to rate their level of agreement with Trump's misstatement and to answer a question about their attitudes toward environmental regulatory policies. The results showed that correction of misstatements make people more accurate regarding those statements and the facts they point to. So, it seems that fact-checking works. However, although people were more factually accurate after a correction, their regulatory preferences and attitudes were indistinguishable from those who did not receive a correction. To put it another way, gaining accuracy on a factual matter does not necessarily affect other related attitudes and policy preferences: participants' attitudes toward environmental regulation remained the same, regardless of their factual accuracy improvement.

These findings suggest that even when our perspective about the world changes, this does not lead us to change our preferences. Republicans and Democrats, or other supporters of two contending political parties, may even agree on the relevant information regarding climate change and still prefer different policies on it. This possibility presupposes that agreeing on all relevant facts and agreeing on how to act on a given issue are two different things. What is such a difference? What do we do when we convey information about how the world is? Is it radically different from what we do when we express information about our own preferences?

In order to measure ideological polarization, we need to know what a population really believes. On the other hand, to measure the type of affective polarization that has to do with the levels of radicalism of a population, we need to know their practical attitudes linked to their level of confidence in the core beliefs of the political group that they identify with. Two people can have different level of confidence in the same core beliefs of an ideological group, which is shown through the things each one is willing to do, their practical attitudes. So, to carry out both tasks successfully, we argue, the difference between the descriptive and the evaluative must be taken into account. As we have seen in the previous chapter, being in a particular dispositional state of mind, such as a belief, depends on the commitments one has, and not just on the commitments one sincerely says one has. Moreover, through our behavior and our statements, be they mental self-ascriptions or statements of another type, one can express one's practical attitudes, that is, one can express information about what can be expected from oneself.

In this chapter, we argue that through the evaluative use of language people express relevant information for knowing what people really believe, but specially for knowing people's affective attitudes, those closely related to their level of radicalism. This distinction between the descriptive and the evaluative, hence, is crucial to the notion of polarization in attitudes that we offer in this dissertation as a result of our reassessment of the concept of affective polarization (chapter 8). In particular, this distinction will enable us to argue that some of the tools commonly used to measure affective polarization actually measure not the feelings that respondents self-ascribe, but their practical attitudes associated with their level of radicalism. If this distinction is ignored, then perhaps it might be the case that someone wants to measure ideological polarization and is actually measuring affective polarization. Or the other way around. In this sense, the distinction between the descriptive and the evaluative will not only allows us to measure polarization more accurately, but also to measure people's affective attitudes, those attitudes with an especial link to action and tied to certain level of radicalism. But before showing all this in more detail (chapter 8), we have to introduce first the distinction between the descriptive and the evaluative from an intuitive point of view and embrace a semantic theory able to account for this distinction and compatible with the view about mental state ascriptions that we have introduced in chapter 6.

We need a semantic theory that accommodates the distinction between the descriptive and the evaluative, and that allows us to explain why the tools that involve an evaluative use of language, such as the tools used to measure affective polarization, actually often measure our attitudes especially linked to action. But not only this. We need to develop ways of measuring polarization that allow us to measure as soon as possible the increase in the level of radicalism in a population. To do so, we need to measure the practical attitudes of the population, those attitudes especially linked to certain courses of action, i.e., what is reasonable to expect from those people. One semantic theory that handles particularly well the task of accommodating the evidence about meaning, that is, our intuition regarding the difference between what we do when we describe our surroundings and what we do when we make an evaluation of our surroundings, is expressivism.

According to expressivism, there is a fundamental difference between our descriptive and evaluative claims. Descriptive claims such as the utterance of the sentence “The laptop is on the table”, when it is used in a particular context, simply expresses our particular belief that the laptop is on the table. That is, one cannot say that one does not believe that the laptop is on the table after making such a claim, even if it turns out that one does not actually have such a belief. Evaluative claims, on the other hand, express not only a particular belief, but something else. For instance, the utterance of the sentence “The laptop on the table is the best one”, when it is used in an evaluative way, not only expresses the particular belief that the laptop on the table is the best one, but also certain attitudes especially linked to action, certain information about what can be expected from the speaker, such as that she would choose that laptop over any other. In this case, it might also turn out that the speaker does not have the belief that the laptop on the table is the best one, but, however, it cannot be the case that the speaker does not have the affective attitudes she has expressed through her evaluative claim. This type of information expressed, the evaluative information, is more connected to the rule one follows and not to the rule one says one follows. Through the evaluative use of language, we argue, people precisely express the kinds of attitudes related to their level of radicalism, and hence it enables us to measure polarization in attitudes. In this sense, this theory helps to meet the desiderata DISANALOGY and INTERVENTION. And, to the extent that this theory enables us to accommodate some evidence regarding the type of information we communicate on different situations, this theory also allows us to go a step further in satisfying the desideratum EVIDENCE. The possibility of error in identifying the rule we actually follow (chapter 6) makes it possible for someone to say that she is describing and yet actually be making an evaluation, but also for someone to say that she is making a positive evaluation and yet actually be making a negative one, and the other way around. All of this is relevant to measure polarization, both the beliefs that people actually have in terms of their contents (ideological polarization) and the affective attitudes that people express to have, linked to their level of radicalism (affective polarization).

This chapter is structured as follows. In section 7.1, we present the descriptive vs. evaluative distinction from an intuitive point of view, and introduce some tests, arguments and considerations that support it. In section 7.2, we introduce expressivism, a semantic theory that can accommodate the evidence regarding this distinction. In section 7.3, we offer our favored noninternalist sort of expressivism, starting from the 'minimal expressivism' and drawing on some of Wittgenstein's insights. We think that having a dispositional mental state and expressing the attitudes and commitments we express through the evaluative use of language are two sides of the same coin. Thus, the view of mental states previously introduced, based on Wittgenstein's philosophy, and the way we conceive expressivism here, as we will see, are not two different theories, but the same one approached from different points. Finally, in section 7.4, we briefly discuss some contextual determinants of the evaluative meaning and the distinction between our judgments and our claims about the rule we think we follow. The information we express through our claims is highly context-dependent.

## 7.1. The descriptive vs. evaluative distinction

Let's start by giving an example of the difference between the descriptive and the evaluative. Take the sentences "Abortion is illegal in this country" and "Abortion is *wrong*". By uttering the first sentence, in the suitable context, we are merely reporting how things are in this country. In particular, we are informing that in this country abortion is prohibited. On the contrary, if we utter the other sentence, in the proper context, we are not reporting how things are in this country, or at least not only that, but something else: we are taking a stance on it, i.e., we are showing our *preferences* and *commitments.* For example, we are saying that we disapprove of abortion and, therefore, that would prefer a country where abortion is illegal and that we are not likely to vote for a pro-choice political party. When we convey information about how things are, we are making a *description.* On the other hand, if the information is not about how the world is, or not just about that, but also about our own perspective, then we are making an *evaluation.*

Intuitively, there seems to be a difference between what we do when we report how things are around us and what we do when we give our opinion or evaluate our surroundings (see Cepollaro et al. 2021; Soria & Stojanovic 2019). In fact, both the findings that correction of misleading information make us more accurate in our factual beliefs without affecting our political preferences (Porter et al. 2019) and the phenomenon of crossed disagreement (Osorio & Villanueva 2019) seem to presuppose the distinction between factual or descriptive information and preferences or evaluative information. Being able to correct our factual beliefs without affecting our political preferences on the same issue presupposes that factual information is distinct, and sometimes independent, from preferences, and therefore it assumes the distinction. A crossed disagreement situation, on the other hand, is based on the possibility that each part displays clear signs of conceiving the disagreement as if the information subjected to discussion were of a different nature, and in this sense it also assumes the distinction.

An evaluation is not always just a matter of uttering some kinds of expressions. Despite the fact that the utterance of terms like 'should' or 'wrong' quite often indicates the presence of an evaluation, it does not necessarily guarantee that the information is evaluative in nature. First, as Wittgenstein points out, we can use certain expressions considered evaluative simply to describe a fact: we can say, for instance, that a person is a *good* pianist, simply meaning that she meets certain criteria and can play pieces of a certain degree of difficulty with a certain degree of dexterity (Wittgenstein 1965), or we can say that abortion is *wrong* simply meaning that in this country abortion is illegal. In these cases, then, both claims can be descriptive ones despite they involving 'evaluative terms'. Second, we can use certain expressions commonly used to evaluate simply for describing a certain standard or norm (Field 2009, 2018). For example, claiming "According to Christian standards, abortion is wrong" is a way of describing Christian standards, that is, a way of reporting one thing that Christian standards sanction. So, by making explicit the standard on which an evaluation is grounded, the apparent evaluation can become a description, because what is said through it is that according to the norm or principle X, certain action or thing is admitted or prohibited. It is important to clarify, though, that the presence of the standard does not necessarily make a claim no longer evaluative, nor the other way around: the fact that the standard is not present does not necessarily make a claim evaluative. Thus, by uttering such things, the speaker does not necessarily say much about how she evaluates Christianity and the pianist: the descriptive vs. evaluative distinction is a matter of sets of language uses rather than sets of terms and expressions. Note, however, that this remark is compatible with the claim that certain sets of expressions frequently indicate evaluation or description due to their widespread use. Henceforth, and for convenience, every time we say that an expression is descriptive or evaluative we assume that it is a descriptive or evaluative use of the expression.

In the previous chapter we have already introduced some tests, provided by Wittgenstein, to try to discern when a statement belongs to the realm of the descriptive and when it is a non-descriptive expression. For example, we saw that when an expression can be classified as a distribution of objects in space and time, or when it can be space-time measured, then quite likely that expression is a descriptive one. In what follows, we will introduce other tests, observations and arguments that support the intuition behind the distinction between the descriptive and the evaluative.

### 7.1.1. The irreducibility of the evaluative

In *A Treatise of Human Nature*, David Hume distinguished passions from reasons, which, according to him, are impressions rather than ideas. Among the direct passions he includes desires, hopes, and fears, which emerge from good or evil, from the pain or pleasure, that we experience (Hume 2007 § 2.1.1.4). Intentional actions arise from direct passions. Thus, passions are characterized by its motivational character. "Reason alone can never be a motive to any action of the will" (Hume 2007 § 413). He argued that morality influences our action and, therefore, cannot come from reasons. In that sense, he distinguished two domains, one of them belonging to reasons, and the other having to do with our motivation to action, to which morals pertains. That's one of the original points of the idea that the motivational character of moral vocabulary comes from a particular kind of mental states: desire-like.

Furthermore, Hume famously pointed out, against moral rationalists, that they make an unremarked and objectionable transition from how things *are* to how things *ought* to be, in part because in doing so they jump from the realm of reasons to the realm of passions. This observation has been widely conceived as the general claim that an evaluative judgment cannot be validly inferred from a set of descriptive premises, and has been called 'Hume's Law'. We take here a more modest interpretation of it, according to which what Hume's remark states is that from a set of descriptive premises cannot be *necessarily* inferred an evaluative conclusion; however, sometimes it can be validly done. The reason is that, for example, we think that one can reach the conclusion that torture is wrong, or that some aids to certain population ought to be promoted by the government –two evaluative claims– after being exposed to some torturing practices or after knowing some data about that population. Consider, for instance, the plot of the movie *Sully*. The main character of this movie, an experimented airplane pilot, and the company he works for, disagree on whether the pilot should have taken another different option than landing the plane in the Hudson River after the engines suddenly shut off. In principle, this is a normative dispute. However, the fact that in several simulations carried out in similar circumstances the pilots did not manage to reach the nearest airport settled the matter in the movie. This does not mean that those facts necessarily will settle the dispute; it simple means that sometimes a fact can settle a normative or evaluative disagreement.

George Edward Moore, in his book *Principia Ethica* (Moore 1993), offered other arguments that go along the lines of showing the distinctiveness of the realm of ethics and the normative. In particular, he argues that moral and normative judgments are *sui generis*, simple and unanalyzable, i.e., they cannot be reduced to, nor implied from, non-moral judgments. The main argument offered by Moore is the well-known "open-question argument". Basically, the argument states that if we try to substitute the predicate "is good" –or any other moral or normative predicate– in the sentence "That action is good" for another allegedly equivalent predicate, like "is pleasant", the resultant sentence "That action is pleasant" leaves open the question whether that action is good. Therefore, the argument goes, being pleasant is not equivalent to being good. The strength of this argument is that for every predicate allegedly equivalent to being good, the question *is it good?* still makes sense after substituting the predicate 'being good' for the allegedly equivalent predicates. Hence, this argument presses in the direction that moral predicates in particular, and the evaluative or normative in general, cannot be reduced to non-normative predicates.

Allan Gibbard offered a more refined and demanding version of Moore's open-question argument, the so-called "What's at issue" test (Gibbard 2003: 23-29). Gibbard proposes to put the question in a form of disagreement between two persons, and analyze whether both claims can be held by the contenders without contradiction. Suppose that two philosophers, Hedda and Désiré, disagree on the meaning of 'good'. Désiré claims that 'good' just means desired, which is the claim to be tested. Hedda, on the other hand, thinks that only pleasure is good. So Hedda claims the sentence "Only pleasure is good", and Désiré rejects it by uttering the sentence "Not only pleasure is good". By uttering these sentences, they clearly disagree: the second sentence contradicts the first one. However, if Désiré tries to express her disagreement by uttering the sentence "Not only pleasure is desired", in which 'good' has been substituted by 'desired', then she will fail, because Hedda can assume this last sentence without contradiction. That is, Hedda can claim that "Only pleasure is good" and that "Not only pleasure is desired" with coherence. Hence, it follows that the sentences "Not only pleasure is good" and "Not only pleasure is desired" do not mean the same thing and, therefore, 'good' is not equivalent to 'desired'. This refined version of the argument makes stronger the claim that the evaluative and the descriptive are two different and distinct domains of language.

### 7.1.2. Action-guidance: Disagreement and other tests

Beyond the arguments mentioned above, there are other issues that point in the direction to the distinction between the descriptive and the evaluative, such as the nature of different types of disagreements and certain linguistic tests that we will review in this section. Both things seem to capture what Gibbard calls 'action-guidance': "To differ over an ought is to disagree about what *to do* or what *to feel* or what *to accept* in some circumstance" (Gibbard 2012: 44, our italics). Evaluative claims seem to reveal an intimate connection to action, i.e., they are constitutively linked to what to do, with a way of living; while descriptive claims seem dead in that sense. Claiming that x is good, or positively evaluating x, requires a commitment or motivation to pursue x if it is possible. If I evaluate something as good, beautiful, tasty, right, etc., I'm giving information about my commitments to act in certain ways. On the other hand, if I describe something as red, flat, rough, dry, etc., I am not mainly conveying information about what courses of action can be expected from me. Consider the following example. If I say that Soto Asa's music is amazing and absorbing, I am communicating, among other things, some of my practical commitments. In particular, if after stating that Soto's music is amazing and absorbing, I refuse to listen to his music on a regular basis and don't flinch at all when his music plays, or even express boredom, then there would be a reason to question whether I really consider it as amazing and absorbing. By contrast, if I say the song "Dra Drari" by Soto was released in 2018, I'm hardly communicating my commitment with one course of action or another; in describing the date of a Soto's song release, I am saying something about the world, I am putting the focus on how things are. In the metaethics literature, this feature that the evaluative exhibits is called action-guidance or practicality (see Gibbard 1990).

One of the characterizing features of evaluative judgments seems to be its gradable and multidimensional nature. As McNally and Stojanovic highlighted, when we evaluate an object, an event, a situation, a person, etc., the evaluation can come in degrees, and we might hinge on different features of the context constituting the 'threshold of applicability', as they call it (McNally & Stojanovic 2017: 21). For instance, when evaluating a Soto's song as amazing and absorbing, we can say that it is more or less amazing and absorbing, and we can give consideration to different dimensions, such as the lyrics, the rhymes, the rhythm, etc. Two tests related to the gradability and multidimensionality characteristics of certain claims are the following. A descriptive claim such as "The song "Dra Drari" was released in 2018" does not admit of degrees −e.g., # The song was very / slightly / more released than another− nor distinguishing different respects playing a role −e.g., # The song was released in every / some respect / with respect to. However, an evaluative claim such as "Soto's song is nice" does admit degrees −e.g., Soto's song was very / slightly / nicer than the previous one– and multidimensionality –e.g., Soto's song was nice in every / some respect / with respect to the lyrics, but not with respect to the rhymes (Cepollaro et al. 2021). Being gradable is a cue of being evaluative, though it does not warrant it.

Soria and Stojanovic have distinguished some additional tests (Soria & Stojanovic 2019). The first one is the *Juxtaposition with 'although'-type connectives* test, based on the action-guidance of the evaluative. The idea is that a complex sentence composed of two sentences connected by the connective 'although', establishing a contrast between a claim and a practical attitude, is correct only when the claim is evaluative. For instance, the claim "Soto's music is amazing and absorbing, although I don't have any intention or plan of promoting or listening it" seems acceptable because the word 'although' marks a contrast assumed by the evaluative content expressed by the antecedent. On the other hand, the sentence "Soto's song is from 2018, although I don't have any intention or plan of promoting or listening it" does not seem quite acceptable, because in this case the antecedent does not express any practical commitment. Another test is the *Lack of epistemic justification*. The idea is that specific evaluative judgments, such as claiming "I've just listened to Soto's music and it is amazing", doesn't admit the question "How do you know?", while descriptive claims such as "Soto's last song lasts 3 minutes" does. Finally, we also have the *Lack of lying potential* test. The idea here is that evaluative claims do not admit the reply "That's a lie", in part because they are a genuine expression of our own perspective. For instance, the claim "This song is amazing" cannot be replied by saying "That's a lie", while a descriptive one such as "This song is the second one of his career" does. Although these tests do not work in every case, taken together they point in the direction that there is a difference between the descriptive and the evaluative.

One powerful way to emphasize the action-guidance of the evaluative is through the kind of disagreement it raises, in contrast to the type of disagreement that can be generated by a descriptive claim. Suppose that after claiming that the song "Dra Drari" by Soto is from 2018, my partner, Ana, disagrees by saying that the song is much older. In that situation, we can settle the dispute perhaps by checking the release date of the song. That is, it seems that we both share the relevant standard, we agree on the relevant information to settle the dispute. Suppose that after doing so, we correctly checked that the song was released on October 5, 2018. Ana, who responds by saying "Well, it seemed like a very long torture to me", has to admit that she was wrong. However, suppose that we disagree not about the song's release date, but about how we evaluate it. Suppose that Ana says that it is a tortuous song similar to how a cicada chirps in the middle of summer. On the contrary, I claim that the song is absolutely gorgeous, in particular because of its addictive spatial sounds. By contrast to the previous disagreement, in this case there is not necessarily a way of settling the dispute that we recognize from the beginning. This is so in part because we are not talking about a property of the song of the kind its duration or the date it was released are. Of course, we are talking about the song, but not just in a descriptive way. Our disagreement shows that we have different attitudes toward the song. That is, we have different ways of living, different standards from which we evaluate it, and part of the information we convey expresses our particular perspective. Of course, when we enter into this kind of disagreement and others, we naturally think that the other part is wrong, and the dispute might be settled. This type of disagreement simply is one in which there is not necessary an agreement from the outset about the things that would settle the dispute. If, in a factual disagreement, we discover that we have different standards, then the disagreement might become meaningless, or it might become about the standard. However, if, in a non-factual disagreement, it is made explicit that we have different standards, the disagreement might neither become meaningless nor become about the standard; it might still continue.

The second kind of disagreement is known as *not-straightforwardly factual disagreement* (Field 2009). As we have seen in section 4.5, within the class of non-descriptive disagreement we can distinguish at least two different types of disagreement, which we have called *normative* and *evaluative* disagreements. Normative disagreements are non-factual disagreements in which the dispute revolves around the standard that should be adopted. For example, if in the last case of disagreement Ana and I turn to discuss about the standard that should be adopted in order to consider the song as a good one, then the resultant disagreement would be normative. However, the dispute can also continue without turning to the standard that should be adopted, but centering on whether the song is good or not. In that second situation, the disagreement would be an evaluative one. The key difference between factual, normative and evaluative disagreements is that in the former type the contenders in dispute share the standards from which to check the truth or falsity of a claim. In the case of a normative disagreement, the parts in dispute do not share the standards from which each one evaluates, and the discussion becomes about the standard that should be adopted. Finally, in a case of evaluative disagreement the parts do not share the standard from which they evaluate, but the discussion does not become about the standard that should be adopted (see Osorio & Villanueva 2019). To be clear, this does not mean that normative and evaluative disagreements cannot be settled, nor that both parties are right. It simply means that, in contrast to factual disagreements, these types of disagreement work differently.

### 7.1.3. The retraction test

Retraction is a movement of language that speakers can make in order to try to take back something they have previously said. According to this general definition, retraction consists in undoing some changes that were previously made in a conversation, regardless of whether they were introduced by a question, an order, an offer, an assertion, etc. (MacFarlane 2014: 108). Thus, one can retract not only something that is now considered false, but also a question or offer, thereby eliminating the obligation of the hearer to respond to it. Note, however, that this general notion of 'retraction' does not require that the speaker changes her mind before retracting. That is, one can withdraw a question simply because one realizes that it has had a different effect than expected, or can withdraw an offer simply because one is tired of waiting for an answer (see Bordonaba & Villanueva forthcoming).

The notion of retraction that has received special attention from the philosophy of language is not the previous notion, but a more demanding one. According to this second notion, a speaker can only retract something that she previously took to be true and now considers false. MacFarlane expresses this idea with what he calls the *Retraction Rule*:

> **Retraction Rule**: An agent in context c2 is required to retract an (unretracted) assertion of p made at c1 if p is not true as used at c1 and assessed from c2. (MacFarlane 2014: 108)

According to the Retraction Rule, hence, retraction is required when a speaker considers from a context c2 that something she previously asserted, in a context c1, is false in the very context c1. For instance, if yesterday (c1) I asserted "It was raining in Granada on August 13", and today (c2) I find out that it did not rain in Granada on August 13, then the Retraction Rule demands that I now (c2) retract what I said yesterday (c1). Therefore, this more demanding sense of retraction requires that a speaker retracts when she now believes that something she previously considered true is false. The speaker must have changed her mind.[1]

Take the following case. Imagine that Antón has studied the literature on differences in intelligence more than anyone else and is prepared to argue coherently, sincerely, and vehemently for equality of intelligence, and in fact has argued the point repeatedly in the past. At the same time, he does have racist spontaneous reactions from time to time. Let's imagine that one day he is talking with his fellow teachers about an Andalusian student with whom he has had some troubles. Someone says that the student is excellent, and Antón says "Andalusian students never submit essays as excellent as students from the north of Spain do". With that utterance, Antón has made an evaluation. In particular, he has reacted in a racist way. Imagine that he suddenly realizes what he has said

[1] There is an interesting debate currently open about whether retraction is a mandatory move. Some authors argue, against MacFarlane, that retraction is not at all mandatory, but optional; one can reject to retract something that was considered true in the past and now considered false without being irrational or insincere, at least in cases of deontic, aesthetic and personal taste predicates (see Marques 2018). Other authors think that the mandatory nature of the retraction depends on each case, specifically on how the case is described (Bordonaba 2017), and that the strength of the demand for retraction depends on whether the statement in question is evaluative or descriptive (Bordonaba & Villanueva forthcoming).

and feels deeply ashamed and regretful. Suppose he tries to retract his comment by saying "No, it is false that Andalusian students never submit essays as excellent as others do". Does he successfully retract his claim with that? We think the answer is no. Of course, he can show his regret. Surely, in such a case, he would wish he hadn't reacted like that, and can claim that what he said is false. But in doing so he cannot completely take back what he has said; he is only showing his regret. In other words, what he regrets is having the *attitude* he had, he is ashamed of having the picture of the world he has, i.e., reacting as he reacts in certain situations, and having promoted certain pernicious stereotypes against Andalusian people. He may try to train his dispositions and change his behavior so he doesn't do it again. But he cannot withdraw what he has already done only by saying that what he said is false. One cannot retract the way one is, that is to say, her attitudes, her commitments, her way of living, simply by saying that what one has said is false; and that's part of the information expressed by an evaluative claim. It is the practical dimension of the evaluative that seems to work differently regarding retraction, because it is essentially tied to our mind, and we cannot retract a claim if we haven't changed our mind: change our mind is not a matter of just saying that we have changed our mind. As argued in previous chapters, saying that one has certain commitments does not necessarily match with the commitments one actually has. The attitudes and practical commitments conveyed through the evaluative use of language are hard to retract, at least in the same way that descriptive claims can be. In that sense, retraction can be used to test whether a particular claim, in a specific context, is evaluative or descriptive.

Of course, the retraction test is not a test that sharply and clearly separates evaluative from descriptive uses for every case; there are cases where the retraction of at least part of the information communicated through an evaluative use of language does not seem particularly problematic. But it seems that retraction works differently in descriptive and evaluative cases. Thus, this points again in the direction of the distinction between the descriptive and the evaluative.

## 7.2. Accommodating the distinction: Expressivism

As we have said at the beginning, a semantic theory that handles particularly well the task of accommodating the difference between the descriptive and the evaluative is expressivism (see, for instance, Frápolli 2019).

Expressivism is the label given to a family of semantic theories that attempt to explain in non-descriptive terms the peculiar meaning and function of the evaluative use of language –it takes as its starting point the idea of discursive pluralism, i.e., that language can be used for different purposes and not just to describe and refer to the world. Thus, one of the core theses of every expressivism is that at least some region of language does not aim at describing how the world is. The motivation behind all or almost all expressivist approaches is twofold. The first motivation is to avoid populating our ontology with entities of a 'spooky' nature. If we say that terms like 'good', in their evaluative use, refer to properties that exist in the world, then we commit ourselves to the existence of such a spooky entity, namely *goodness*. This is what Price has called the 'placement problem' (Price 2011).[2] The second motivation is to capture and explain the close connection to action of the evaluative language. To say that a certain thing is good is, as we have seen, to show that we are motivated to pursue that thing. There is a motivational component, an attitude, that is shown with these types of claims. If terms like 'good' refer to properties in the world, then it is difficult to explain this motivational component. Thus, the expressivist avoids populating our ontology with 'weird' entities and opens the door to an explanation of the motivational component of the evaluative use of language, by denying that the function it accomplishes is referential. Instead, the function that this type of use of language fulfills is expressive. What is exactly expressed through it depends on the particular expressivist theory.[3]

[2] See Navarro-Laespada forthcoming for an interesting and recent discussion about the placement problem.

[3] There is an ongoing debate between global and local expressivists (see Frápolli 2019). Global expressivists hold that all regions of language accomplish a non-descriptive function. Local expressivists, on the other hand, hold that just some regions of language fulfill that function, but other regions do describe how the world is. We will not spend much time on this debate here, because it is

One of the main features of expressivism is what Gibbard calls "the oblique strategy" or "the oblique analysis", consisting in explaining the meaning of an expression not by focusing on its truth-conditions, but on the kind of mental state we express through it.

> The expressivist now turns to oblique analysis: we elucidate the concepts of ought, meaning, and mental content by saying what it is to judge or believe that a person ought to do something, or that he means such-and-such or that he is thinking that such-and-such. (Gibbard 2003: 193)

The meaning of an expression or term is explained by virtue of "what states of mind the term is used to express" (Gibbard 2003: 5-6; see also Wedgwood 2007: 35). In particular, the evaluative involves the expression of mental states especially tied to action. Standardly, this distinction has been conceived as one between cognitive vs. conative mental states, where mental states are understood as internal things. When the utterance of a sentence expresses a cognitive, belief-like –i.e., representational– mental state, then the meaning conveyed is descriptive. On the other hand, if the utterance of a sentence expresses a noncognitive, desire-like mental state, then the meaning conveyed is evaluative. This distinction finds its sense in that conative, desire-like mental states are traditionally associated with action, while cognitive, belief-like representational mental states are associated with how the world is.

The earliest varieties of expressivism are often attributed to Ogden and Richards (Ogden & Richards 1923), Ayer (Ayer 2001), Stevenson (Stevenson 1937), and Hare

not central to our purpose. However, we would like to say something briefly about it. We agree with globalists that there is no region of language that acquires its meaning from how the brute reality is, i.e., from extralinguistic facts; meaning is a normative issue, and there are no extralinguistic facts determining the meaning of any statement, as Wittgenstein and Kripke famously showed. However, we disagree with some globalists on the claim that language works in a homogeneous way. Also following Wittgenstein, we think that language follows different purposes, and one of them is descriptive. In fact, some authors have recently pointed out that local and global expressivists are not in conflict (Simpson 2020). The alleged incompatibility between them can be dissolved if it is understood, for instance, in terms of metasemantic considerations vs. semantic insights.

(Hare 1952), respectively. Despite the differences between them, the canonical interpretation of what is known today as classical expressivism, a kind of proto-expressivism, is based mainly on some of Ayer's ideas. According to Ayer, ethical statements, in contrast to empirical claims, have emotional meaning. In particular, these statements have at least two types of possible meanings. The first of them is relative to the morality of a group of people. According to this first sense, to say that something is right is to say that it is among the things permitted by a set of norms, which is to make a description, and therefore is similar to what we do with empirical statements. The second of the two possible senses is the truly ethical one. According to this second sense, ethical statements express the speaker's subjective emotions, specifically her approval or disapproval towards something. In this sense, saying that something is wrong is similar to booing it, and saying that something is right is like saying hurray for it!

The canonical interpretation of this position claims that according to classical expressivism, evaluative statements do not express propositions at all and, therefore, do not have truth-conditions. This interpretation has given rise to what has been taken to be the central problem facing expressivism, the so-called Frege-Geach Problem (Geach 1960, see Schroeder 2008). According to the classical formulation of this problem, if ethical statements do not express propositions capable of being declared true or false, then it is difficult to explain what happens in our daily reasoning of the type 'if it is true that p, then q', where p is an ethical statement. The antecedent of the previous conditional has to express something true or false for the full reasoning making sense. If, according to classical expressivism, the claims of ethics are neither true nor false, then these kinds of arguments would be meaningless. But it seems that we usually reason in this way.

However, Ayer did not claim that ethical statements do not express propositions and therefore don't have truth-conditions, but that ethical expressions do not make any contribution to the proposition expressed (Ayer 2001: 110; see Frápolli & Villanueva 2012). If the sentence "The boy has stolen the ball" expresses a truth-apt proposition in Ayer's framework, then the sentence "It is wrong that the boy has stolen the ball" also expresses a truth-apt proposition; the ethical expression 'It is wrong' simply adds nothing to its factual meaning. Hence, the canonical interpretation according to which moral expressions block the possibility of being truth-apt is erroneous. Contemporary hybrid expressivism takes its cue from this second interpretation to construe their proposals, according to which evaluative claims express propositions and are truth-apt.

Since the emergence of these early expressivist theories, the idea that beliefs have no link to action has remained among expressivists. However, all assertions, even the evaluative ones, express beliefs, because asserting p is conceptually tied with acquiring the commitment that one believes that p, and also that one thinks that p is true. In other words, a speaker cannot assert p and then deny that she believes that p, or that p is true, without being incoherent. So, all assertions necessarily involve the expression of a belief. Moreover, as we have said, there may be beliefs with a strong link to action, such as the belief that the government is responsible for the pandemic, or that fascism is advancing dangerously, and conceiving mental states in descriptive terms is problematic, as we have seen in chapter 5. Therefore, we will propose that it is better to simply focus on what it is to judge or believe that p when p is descriptive vs. when it is evaluative, and not in terms of cognitive vs. noncognitive mental states –or to reinterpret the cognitive vs. conative distinction in these terms. To judge or assert that p, when p is evaluative, is to express, in addition to the belief that p, other mental states *especially* linked to action, attitudes. On the other hand, when p is descriptive, to judge or assert that p is to express just the belief that p, a mental state not especially linked to action.

Finally, we want to stress another idea that, although it is not held by all positions deemed as expressivists, we consider a key piece of expressivism. This idea is the contrast between *expressing* a mental state and *saying* that one is in it (see Gibbard 2003: 76). That's what we have called the DISANALOGY requirement. Expressivism not only distinguishes between descriptive and non-descriptive claims, but also between self-reports of the rule we follow, and expressions of the rule we actually follow. In other words, if the evaluative accomplishes the function of expressing our practical state of mind, then the state of mind in which we say we are and the mental state we express to be in may not fit, and we can express another different mental state. On the one hand, we have evaluative expressions, like "It should be illegal to demonstrate against wearing masks", that express our mental state. On the other hand, we have self-reports of mental states, like "I believe that it should be illegal to demonstrate against wearing masks". Despite the fact that in claiming the former I imply the latter, they have different meanings because one can agree on that it should be illegal to demonstrate against wearing masks but disagree on whether I believe so. That is, I can say that I believe such a thing, but express through my linguistic and nonlinguistic behavior that I don't actually believe so, and even that I have other affective attitudes. If I make an evaluation, then I express the mental state in which I am, my attitudes. But if I report the mental state in which I am, I am not necessarily in that mental state. To be clear: our claims, be they mental self-reports or claims of another type, can show that we don't follow the rule we sincerely say we follow, and can also express that we follow other different rules, that is, can express our affective attitudes. These possibilities, we think, are essential to a viable expressivism, which can only be accommodated by *noninternalist* versions of expressivism.

### 7.2.1. Contemporary Expressivism

A contemporary way of explaining the distinction between the descriptive and the evaluative is based on dynamic semantics: *dynamic expressivism*. Dynamic semantics analyze meaning in virtue of the actions performed with a piece of language in a particular context. In particular, dynamic semantics treats meaning as individuated by the changes it effects in a given conversational context. The origins of this framework can be traced back to the work of Groenendijk and Stokhof (Groenendijk & Stokhof 1991). Within this framework, the conveyed information is conceived as an updating step allowing to replace a previous information state by a new one. Information states are called contexts. Thus, the meaning of the utterance of a sentence in a particular context is determined by making explicit the role it plays in updating the context in which the sentence was uttered.

One crucial point is that an utterance, within this framework, can mainly play two different roles. On the one hand, it can help to situate the actual world in a region of the logical space by eliminating those possible worlds of that region which are incompatible with the truth of the utterance of a sentence in that context. On the other hand, it can serve not to update the place of the actual world in a particular region of the logical space by eliminating incompatible possible worlds, but to make partitions in the region of the logical space and rank them. When an utterance performs the first function, it conveys locational information; when it plays the other role, the information conveyed is orientational (Charlow 2014; Lewis 1979; Soria 2019). As Yalcin puts it, "normative discourse is distinctive in respect of its dynamic effect on the state of the conversation" (Yalcin 2018: 400).[4]

The idea is that the evaluative is explained as mainly conveying orientational information rather than locational one. When I said to my partner "The song "Dra Drari" by Soto is from 2018" I am proposing to update our set of shared knowledge about how is the actual world, i.e., the common ground (Stalnaker 1978). In particular, I am proposing to my partner to eliminate every possible world of the common ground where the song was released on another date. In that sense, the information provided is descriptive. On the other hand, when I said to my partner "Soto's music is amazing and absorbing", despite assuming that the actual world is one where Soto's music exists –something already included in the common ground, I am not saying anything about the world, i.e., the set of possible world of the common ground cannot be updated in the sense of eliminating those incompatible with the new information, because "is amazing and absorbing" does not say a word about the actual world. With it, I'm just proposing to order those worlds in a particular way, maybe ranking them by virtue of how preferable they are, or making partitions in the logical space.

Dynamic expressivism is to some extent built upon another very influential contemporary expressivism, Gibbard's hybrid position (Gibbard 1990, 2003, 2012). In his book *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Gibbard 1990), Gibbard offers an expressivist proposal for attributions of rationality, e.g.,

[4] The possible world framework can be understood as a representation, as a representational system. But it can also be understood in another way, for example simply as certain assumptions that allow us to explain the difference between the descriptive and the evaluative in terms of the locational or orientational information we communicate through a claim.

"X is rational", which he takes as the paradigmatic normative claim. According to him, normative claims express a complex state of "norm-acceptance". In particular, it expresses a factual belief plus a normative state of accepting a system of norms N. For instance, the sentence "X is rational" expresses the factual belief that X is permitted by N and the normative state of acceptance of the system of norms N. Thus, normative claims express states of norm-acceptance, while descriptive claims express just states about how the world is.

Later, Gibbard develops this idea and applies it to sentences of the type 'I ought to pack', offering a way to model the content of declarative sentences that can account for normative claims. In doing so, he keeps the idea that by uttering a sentence we express a mixed or hybrid state of mind, i.e., a mental state with a dual direction of fit –partly world-to-mind, partly mind-to-world. But he adds that the contents of declarative sentences can be modeled as sets of world-hyperplan pairs, in particular as the set of world-hyperplan pairs where the content expressed by a particular sentence is true (Gibbard 2003: 58). This way of representing the contents of declarative sentences hinge on the traditional picture of propositions as sets of possible worlds. In particular, Gibbard's account draws on a model of mental states given by Lewis, which represents beliefs as a set of centered worlds (Lewis 1979). According to this picture, a proposition p is the set of possible worlds, in the logical space, where p is true. For instance, the proposition *the laptop is on the table* comprises the set of possible worlds where it is the case. The second component of Gibbard's pairs, i.e., the hyperplan, is construed on a similar way as propositions are conceived here: as a set of decisions about what to do. In particular, a hyperplan is a plan where an agent has maximally decided what to do in every possible circumstance. Thus, if a possible world is a maximally determined set of facts, a hyperplan is a set of decisions about what to do for all conceivable contingencies.

From this sort of expressivism, the content of a declarative sentence is represented as the pair (w, h), where w is a set of possible worlds, and h is a hyperplan. If the truth-value of the content of a sentence is sensitive just to the first element, then it is descriptive. By contrast, if the truth-value of the content of a sentence is sensitive to the second member of the pair, then it is evaluative. For instance, the sentence “The boy has stolen the ball” is true in some worlds and false in others, but its truth-value does not depend on what the hyperplan is; the sentence will be true in some worlds and false in others no matter what has been decided with respect to every conceivable contingency. Hence, this sentence is descriptive: its truth-value affects only to w. However, the sentence “Stealing is wrong” may be true according to some hyperplans and false according to others, but its truth-value does not depend on how the world is; the sentence is true according to some decisions about what to do and false according to others. Hence, this sentence is evaluative: its truth-value is at least sensitive to h. Note that in the case of the evaluative we have said that a content can be deemed evaluative if it is at least sensitive to h. This claim leaves open the possibility that the truth-value of certain evaluative sentences were not only sensitive to h, but also to w.

Some authors have applied the framework developed by Gibbard to knowledge attribution, giving rise a sort of expressivism known as epistemic expressivism (Chrisman 2007) or evaluativism (Field 2009, 2018). According to Chrisman’s epistemic expressivism, claims of the form ’S knows that p’ express a complex state of mind consisting of the belief that S is entitled by norms e to her true belief that p, and the acceptance of the epistemic norms e (Chrisman 2007: 241). Field’s evaluativism is quite similar, but he introduces some elements from the kind of relativism defended by MacFarlane (MacFarlane 2014), i.e., assessor relativism.[5] According to Field, to attribute knowledge to someone is to make an evaluation, and evaluations are relativized to a standard from which we make the evaluation. The claim of the sentence ’S knows that p’ expresses a proposition that is incomplete with regard to assessor’s norms –assessor’s preferences and policies. On the other hand, the truth of a descriptive claim is just determined by the world component. This is in line with the interpretation of Ayer’s theory according to which, although ethical expressions do not make any contribution to the proposition expressed, evaluative claims express propositions. Something must be a proposition in order to be taken as an evaluation. If there is no proposition, then there is no evaluation. As we will see in the following section, this is

[5] See (Frápolli & Villanueva 2015) for a discussion of the differences between MacFarlane’s relativism and expressivism.

one of the intuitions behind minimal expressivism. Thus, according to these expressivisms, every claim involves a complex state of mind that comprises a factual and a normative component. When the truth of a claim is relative to the factual component, i.e., the factual belief, then the claim is descriptive. When the truth of a claim is relative to the preferences or policies part, the claim is evaluative. That's the main reason why this sort of expressivism is called hybrid.

A first concern with such position is related to the descriptivist nature of the first type of mental state expressed by an evaluative claim. According to this position, an evaluative claim expresses a complex mental state, composed by a factual belief plus a normative state. The factual belief is that certain fact is the case. As Fields notes (Field 2009, 2018), whether something is in accordance with a standard is a statement whose truth is factually determined. However, as we have seen in the two previous chapters, it is highly problematic to assert that belief ascriptions accomplish a descriptive function (see also Frápolli & Villanueva 2018). Moreover, the idea expressed by the expression 'factual belief', i.e., that it is a *fact* that a certain action is permitted by a set of norms, is highly controversial, for similar reasons. Wittgenstein and Kripke draw our attention to the fact that every action or expression, under a suitable interpretation, can accord with a given rule. Thus, even in the descriptive region of language it is problematic to say that there is an external fact that determines the meaning of an expression. Meaning, like dispositional mental state ascriptions, is a normative issue: there are no extra linguistic facts fixing the meaning of an expression nor determining that an action accords with a rule.

Second, these family of contemporary expressivisms are compatible with the idea that mental states are internal things. We think that Gibbard's and Field's proposals can be read as noninternalist expressivisms, where being in a mental state, as we saw in the previous chapter, amounts to being someone from whom certain verbal and nonverbal courses of action can be expected (see, for instance, Gibbard 2003: 77). But this is controversial. As we have seen, Gibbard recognizes the difference between expressing a mental state by claiming, for instance, that I'm planning to pack, and saying that one is in a mental state by claiming, for instance, that I believe I'm planning to pack. However, he also says that "these two states of mind go together, except in *weird cases*" (Gibbard 2003: 76-77, our italics). As we have seen throughout the previous chapters, being in a mental state different from the mental state one sincerely claims to be in is not as strange as it might appear. This, together with the commitment that whether a claim is correct according to a rule is factually determined, can lead us to suspect that Gibbard has an internal conception of mental states. We need a position that accommodates the distinction between the descriptive and the evaluative and avoid these problems.

## 7.3. Noninternalist Expressivism

As we have seen, expressivism, in explaining the evaluative in psychological terms, seems to commit to a picture of the mental that is incompatible with the picture we have introduced in chapter 6. In this section we attempt to sketch a kind of expressivism free from internalist and descriptivist commitments regarding mental states and meaning (Sambrotta & García-Jorge 2018), i.e., a position holding that (i) mental states are not things, neither internal nor external, (ii) mental state ascriptions do not have a descriptive function, and (iii) there is possibility of error in self-ascribing a state of mind and the rule one thinks one follows. To do so, first, it is worth to wonder what are the minimal conditions that a position should meet to count as an expressivist one.

According to Frápolli and Villanueva (Frápolli & Villanueva 2012; Frápolli & Villanueva 2018), the minimal conditions a position must meet to be an expressivist one are:

> **High-Order-Functions (HOF)**: Certain predicables do not take simple objects within their scope, but complexes of objects and properties. These predicables are 'second-order' ones, or functions of propositions.
>
> **Non-Descriptivist (ND)**: Second-order predicables do not describe the world.

> **Truth-Conditional-Irrelevance (TCI)**: Second-order predicables do not modify the truth-conditions of expressions within their scope.

They call this position 'Minimal expressivism'. This way of characterizing expressivism attempts to grasp the distinction between the descriptive and the evaluative in terms of logico-syntactic characteristics. In particular, they claim that certain predicables can take a proposition as its argument –e.g., the predicable 'is good' can take as an argument the sentence 'Mary has finished her work', generating the sentence 'it is good that Mary has finished her work'; on the contrary, the predicable 'is tall' does not exhibit the same logico-syntactic feature, and this possibility is what characterizes the expressions on which expressivism focuses: "the expressions targeted by the expressivist *can* occur as typical functions of propositions" (Frápolli & Villanueva 2012: 472). We agree with this way of conceiving the minimal conditions of expressivism. However, although we agree that every expression that can function as a second-order predicable has the capacity to fulfill a non-descriptive expressive function, we think that the feature of being able to occur as a second-order predicable is not a feature exhibited by all expressions targeted by the expressivist. Expressions of personal taste like 'it is tasty', for instance, do not seem to have the capacity of functioning as second-order predicables. The expression 'It is tasty that ...' does not simply have any felicity uses. Moreover, there seem to be other expressions, such as pejorative terms or dogwhistles, that interest the expressivist and that do not seem to admit the test of second-order predicables. In that sense, we think that HOF does not exhaust the set of expressions that interest the expressivist analysis. Thus, the way we see the negative side of expressivism is as follows. Many expressions serve a non-descriptive evaluative function. A wide set of these expressions exhibit a peculiar logical-syntactic feature, namely: they can function as second-order predicables. However, this feature does not exhaust the set of uses that fulfill an evaluative function.

We start from this sort of relaxed minimal expressivism to try to offer a positive proposal, that is, a proposal about the function the evaluative uses of language fulfill. According to the type of position that we will try to outline here, the difference between the descriptive and the evaluative is built upon the different practices in which we can engage, i.e., different conceptual links with different courses of action tied to them. In that sense, this theory assumes what Frápolli and Villanueva call the *organic model* –in contrast to the *building-block model*: the model that gives prominence to the proposition or judgment as the basic unit of meaning (Frápolli & Villanueva 2015; Frápolli & Villanueva 2016; see also Frápolli 2019). As they put it, the building-block model and the organic model "can be set apart by taking into consideration whether they give prominence to the principle of compositionality over the principle of context, or the other way around" (Frápolli & Villanueva 2015: 1). The idea is that in evaluative practices, instead of conveying information about how things are, our claims communicate information about our own world-view, i.e., our preferences and attitudes. That is, they indicate *circumstances of evaluation*: they express or make explicit certain particular inferential links and actions. The main goal of descriptive practices, on the other hand, is to provide information about how this world is. It is in that sense in which evaluative claims express speaker's mental states *especially* linked to action: they make explicit a way of living, conceptually articulated. To better grasp this idea it may be useful to say something more about the two things we can express through our claims.

One the one hand, by asserting p, the speaker is expressing her belief that p, which means that she cannot deny that she believes that p without being incoherent. In that sense, every time someone asserts that p, she expresses her belief that p, her commitment to not deny that she believes that p. Note, however, that this logical link does not depend on the nature of what is asserted: If I assert that the laptop is on the table –a descriptive claim, then I cannot say that I don't believe that the laptop is on the table; but if I assert that stealing is wrong –an evaluative claim, then I cannot say that I don't believe that stealing is wrong too. Therefore, it is clear that every assertion expresses a belief, in this sense. But, crucially, this type of belief is not especially tied to action, it does not make explicit the speaker's way of living. If someone performs some actions that are conceptually incompatible with having such a belief, or if the contextual information available contradicts the logic of having such a belief, then one might not believe that p despite sincerely asserting p. That's why asserting p, or self-ascribing the belief that p, does not guarantee that the speaker believes that p. The belief expressed in this level is linked to action, but not especially.

On the other hand, evaluative claims express something else, information about the speaker's attitudes, preferences and stances: her way of living. Speaker's evaluations locate her in a place of a socio-normative space from which many things can be truly predicated of her. And, in contrast to the particular belief that p expressed through the assertion of p, the affective attitudes expressed are conceptually incompatible with error. That is, it cannot be the case that someone does not have the attitudes especially tied to action that she expresses to have through her evaluative claim without being irrational or incompetent. To be clear, it could be the case that someone expresses an affective attitude and her subsequent behavior does not correspond to the attitude expressed, but then, the speaker will be necessarily considered as irrational or incompetent, and not just as someone who has said something false. That's one reason why affective attitudes are closely connected to action. Thus, since a mental self-ascription can play an evaluative function, it could be that someone self-ascribe the belief that p and that, through her claim, she is expressing not the belief that p, but her practical attitudes; she is presenting herself as someone from whom certain courses of action can be expected, as in the Obama's example introduced in chapter 6: in his speech, Obama's belief self-ascriptions don't merely express those beliefs, but her practical commitments beyond those beliefs. Being able to differentiate among these things is an advantage of a *noninternalist expressivism*.

One might think that the information conveyed by an evaluative claim is close to the information conveyed by a self-ascription of a mental state in the sense that both are nondescriptive in nature: none points to a distribution of objects in space and time. Nonetheless, as we have said, we must be careful here. It seems that we can be wrong in self-ascribing a mental state, and it does not seem particularly problematic to say that it is false that I was in the state of mind that I claimed to be. On the contrary, when we express information about our mindset through an evaluative claim, we cannot be wrong in the same sense; if our subsequent behavior is not in accordance with the attitudes expressed, then we will be considered irrational or incompetent. The thing is that we can also evaluate through a mental self-ascription. Our mental self-ascription can be false not only in the sense that we are not in the mental state we say we are. Also, we can express our practical commitments through them, as in the Obama's example.

Maybe, the difference we are trying to highlight here can be clarified by pointing to the difference between our judgments and our claims about the rule we think we follow. Note that, despite claiming that stealing is wrong and claiming that it is wrong that Julian stole that thing are both apparently non-descriptive claims, there could be a difference between them. In the first situation, by claiming so maybe I am just saying which is the rule I follow, a very general rule according to which there is no situation in which stealing is not wrong. In that sense, this claim can be seen just as a claim about the rules I think I follow. On the contrary, in the second situation maybe I am making an evaluation and, in that sense, I am not saying which is the mental state I am in, but expressing or showing the rule I actually follow. The idea is that through our evaluative judgments we more frequently show or express the rule we follow, our world-picture. Claims that are apparently non-descriptive, when they are sufficiently general, are more likely to be just claims about the rule we think we follow, while when they are made on specific situations, they are more likely to express or show the state of mind in which one is, be it a particular belief or other attitudes especially linked to action in the sense specified above. But this is not always the case. For example, claiming that all politicians are dishonest, in a particular context, might express some practical attitudes. We will say something more about this in section 7.4.

In sum, through the evaluative use of language, we are not only committed to the idea that by asserting p we cannot deny that we believe that p, but also with the idea that, when they are evaluations, we express our world-view, we express information about the courses of action that can be expected from us, our practical commitments. Only noninternalist versions of expressivism can accommodate the possibility of error to identify our own commitments.

### 7.3.1. Wittgenstein and the evaluative

In a letter to Russell, Wittgenstein observed that “the main point [of the *Tractatus*] is the theory of what can be expressed by propositions −i.e., by language . . . and what cannot be expressed by propositions, but only shown; which, I believe, is the cardinal problem of philosophy” (Stern 1995: 69-70). This distinction between what can be said by propositions and what can only be shown is essential to *Tractatus*. Recall that in the *Tractatus*, what can be said is the set of bipolar propositions that represent states of affairs. Statements about ethics and logic, as well as ascriptions of meaning and beliefs, on the contrary, belong to the realm of what can only be shown. With these statements we do not make movements like those we do with descriptions, because we do not follow the rules of the system to represent a state of affairs, but we point directly to the rules that constitute the system itself. That’s the reason why Wittgenstein says that they do not express proposition, but pseudo-propositions.

As we have seen, this distinction is what supports the notion of description that remains constant throughout Wittgenstein’s work. In his mature stage, Wittgenstein calls ’grammatical propositions’ and ’logical propositions’ what were previously deemed pseudo-propositions. But the idea is similar. Sentences representing distributions of objects in space and time express empirical propositions. Sentences pointing to the rules that govern a language game express grammatical or logical propositions. In that sense, the difference between what can be said and what can just be shown survives until his mature thought.

The crucial point here is that what can be described or said and what can only be shown belong to completely different practices. What can be described exhibits certain conceptual peculiarities. Among them it is, as we have seen, their conceptual connection with predicates such as ‘having spatio-temporal location’, their inability to function as a second-order predicables, and the particular type of disagreements that they generate, namely, disagreements in which both parties recognize a priori the fact that would settle the dispute. It could be said that descriptive practices are characterized by the fact that the rules we follow are widely shared. In other words, making a description is like reporting what happened in a football game to someone who knows well the rules of football. For example, if in such situation we state "Player number 9 passed the ball to player number 7", the information that we communicate is about something that has happened in the match, and that is in accordance with the rules of football. If it turns out that player number 9 passed the ball with her hands, then our interlocutor could deny that this was a pass, and correctly warn that it was an illegal move. Since we both agree on what the rules of football are, we would recognize which of us is wrong. However, if instead of claiming such a thing we say something like "Player number 9 should be able to pass the ball with her hands to her teammates", then we are no longer reporting something that happened in the match, but talking about the rules of football themselves, specifically about how they should be.

In both descriptive and evaluative practices, the meaning communicated through a statement depends on its inferential links, that is, on what follows from the statement in that context. However, both practices are subject to different conceptual links, among other things because they serve different purposes. If I say "Player number 9 passed the ball to player number 7", the hearer does not acquire any especial information about what can be expected from me, that is, about how I see the world. However, if I say "Player number 9 should be able to pass the ball with her hands to her teammates", my interlocutor does acquire an especial piece of information about me. The set of conceptual links with which I engage, and their connection to action, is broader than in the descriptive case. The rules I express to be committed to are not necessarily followed by all people. In other words, the practices supporting what is deemed as good, desirable, and other evaluative attributions do not exhibit a presumption of communality, while the practices supporting the descriptive are widely shared. That's the reason why my interlocutor and I would recognize which of us is right or wrong if we disagree on whether one player has passed the ball to another, and why the disagreement will turn out meaningless if we discover that we have different standards, or it will turn out about the standard. However, this is not the case in non-factual disagreements: even if it is explicit that we have different standards, the disagreement can still be about the same thing. The stability of the shared practices in a form of life enables us to explain the difference between the descriptive and the evaluative.

Of course, we can say that someone who makes an evaluation is wrong. But here ‘being wrong’ works differently than in the descriptive realm, at least in the sense that there is no necessarily an initial shared commonality to which appeal in order to settle the disagreement. It means that from my way of living, from my standard or background, that person is wrong, that is, what she says is false considered from my standards. But maybe there are other ways of living, other standards, from which that same evaluative statement is taken as true, and I’m aware of that: I recognize the possibility that I’m in a mistake. In the descriptive realm, however, there is no room for this possibility. There are no standards within our form of life from which, for instance, to believe that water does not boil at 100 degrees Celsius is to believe something true. Because in our form of life, that water boils at 100 degrees is a hard rock, that is, the practices in which this statement is dependent are widely shared, and many other of our practices rest on it.

Let us say something more about the different practices supporting the descriptive and the evaluative to try to be as clear as possible. According to Wittgenstein, the justification of our picture of the world always comes to an end, that is, there is a moment where *our spade turns on bedrock*. This is so because the bedrock, our background or standard, is just supported by our acting. In other words, our background is the set of things we have decided not to cast doubts about, which is supported by our way of behaving as a community and serves as support for other moves. In that sense, our picture of the world is groundless, and the justification for it always comes to an end. In addition, our picture of the world is our frame of reference, that is, the background from which we distinguish between what is true and what is false.

The thing is that in a form of life, which is a set of practices, there are subsets of those practices shared by all the community, and other subsets just shared by a part of the community.[6] When, in a disagreement, the relevant set of practices

[6] In fact, there are at least two options here. One is to think that if a form of life is a set of practices, then the sub-communities related to different practices belong to different forms of life. In this way, in a community there would be different forms of life. The other option is that all identifiable practices in a given language community make up a unique form of life, and within that form of life distinctive sub-communities with different ways of living can be distinguished. I prefer

are shared by both parts, the disagreement necessarily will be factual, i.e., it will necessarily be a disagreement in which both parties recognize from the beginning which are the facts that would settle the disagreement. On the other hand, when in a disagreement the relevant set of practices are not shared by the parts, the disagreement not necessarily will be settled by appealing to facts, and wouldn't become meaningless nor would become about the standard that should be adopted if it is made explicit that both parts have different standards.

Thus, the descriptive is supported by the practices we share, while the evaluative is supported by the practices not shared by all the community, but just by a part. In that sense, evaluations give information about our own way of living, i.e., about the practices in which we are engaged, and for that reason they convey information about the courses of action that can be reasonably expected from us, that is, our particular picture of the world. Lynch calls 'convictions' the set of beliefs supported by the practices not shared by all the community, and emphasizes its action-guidance component; they are commitments to action. As Lynch puts it, "we see other people's convictions as revealing who they are, and they in turn look at our convictions in the same way" (Lynch 2019: 60). Stanley uses 'ideological beliefs' for something similar (Stanley 2015). What Lynch calls convictions and Stanley ideological beliefs is what we have called, more broadly and following Wittgenstein, dispositional mental states especially linked to action: the set of conceptual commitments especially linked to action that reflect our picture of the world and, in that sense, who we are. This is the way in which evaluative uses of language express our mental states. And, of course, there are mental states in which you are more or less confident, depending on how involved you are in the practices that sustain it, and how central they are in your daily life. Beliefs whose contents exhibit a high degree of belief are usually those that are essential to our

this second option because, from the point of view of meaning, the first option would lead to say that within the same linguistic community, sub-communities that follow different practices use words with different meanings, since the meaning of expressions rests on how they are used. The second option, however, allows us to say that the meaning of expressions is determined by all the practices in which those expressions can appear and have appeared within a linguistic community, thus allowing to explain, for example, the mechanism under which dogwhistles operate, and avoiding to state that disputes between people with different standards are always metalinguistic disputes.

identity. We show our level of confidence in those beliefs through our attitudes.

Let's take an example to see all this more clearly. To claim that it is wrong that Juan has boasted of his achievements, an evaluative claim, could be the result of assuming that "You never show off what you get", whose justification ends in a way of living. In this sense, the truth of "It is wrong that Juan has boasted of his achievements" is relative to such a way of living, a particular standard that depends on a way of living. And this kind of assumption belongs to the realm of what can only be shown because, as Wittgenstein says, in trying to justify it, for every reason we will always find a counter-reason (Wittgenstein 1980a). Let's see why. Someone could live in such a way that one of her assumptions is "One only shows off what one has earned with effort". For this person, then, the statement "It is wrong that John has boasted of his achievements" could be false if Juan reached his achievements with effort. And this person, therefore, could always find a counter-reason for the proposition "You never show off what you get". In this way, making the evaluative claim that it is wrong that Juan has boasted of his achievements expresses the speaker's picture of the world, i.e., her way of living, her practical mental states.

## 7.4. Evaluative meaning: Contextual factors and judgments

Let's take a look back and see what we have done so far in this chapter. First, we have introduced the distinction between the descriptive and the evaluative, a distinction between different uses of language. This distinction is important here because it enables us to measure polarization while avoiding self-deception and in general those cases where people don't exhibit first-person authority regarding their mental self-ascriptions. But also, this distinction is important because it enables us to measure the attitudes especially linked to action that we express through the evaluative use of language. Hence, this allows us to fine-tune our tools in order to measure both ideological polarization and affective polarization. Second, we have introduced a group of tests and arguments that support the distinction between the descriptive and the evaluative. Third, we have introduced expressivism, a theory that seems able to explain the difference and therefore to accommodate this evidence. Expressivism typically explains the difference in psychological terms, that is, by appealing to the difference in the types of mental states we express in both cases. However, for the purposes of this dissertation, it is crucial that this theory can be compatible with the normative approach that we have introduced in chapter 6, a view characterized by recognizing the possibility of error in ascribing mental states. For that reason, we have introduced minimal expressivism, and have discussed how an expressivism inspired by the Wittgensteinian's remarks we have pointed out in the previous chapter would look like.

Although we have presented different tests, arguments and ways of explaining and discerning between descriptive and evaluative uses of language, it is important to note that evaluative meaning is highly dependent on contextual factors and, in that sense, not easy to determine. As we have said at the beginning of the chapter, the evaluative and descriptive distinction is a matter of uses of language rather than of a set of expressions and terms. However, as we have also said, some terms and expressions are more frequently associated with a category because they are mostly used to convey information of that nature. For instance, the expressions "It should be that…" and "It is great that…" are normally used to evaluate and, in that sense, they are associated with the evaluative. Expressions like "There are 10 people in the room", on the other hand, are often used to describe. Based on this, Pew Research Center conducted an experiment in which it asked participants to categorize ten sentences as descriptive or relative to opinion, among which allegedly there were five descriptive sentences and five opinion statements (Gottfried & Grieco 2018). Only the 26% accurately classified the five descriptive sentences, and only the 35% did it well in classifying the five opinion statements. So, as this finding shows, we are very bad in performing the task of distinguishing between the descriptive and the evaluative. But things are even worse. Note that this empirical study was carried out with sentences taken in isolation, where their descriptive or opinion character is determined by the wide use of the sentence, which facilitates the task of identifying the type of information communicated. In everyday and more natural situations of communication, the task of identifying the type of information that someone communicates when saying something is even more complex. What does the meaning we communicate depend on?

As advanced in section 3.4.1, we have conducted an empirical research to test the effects of several contextual factors in considering the utterance of a same sentence, that appears to be descriptive, as offensive (Almagro et al. forthcoming). In two studies, we tested the role of the speaker identity, both in terms of membership (pertaining or not pertaining to the group of people she talks about) and social status (high vs. low social status), the role of speaker intent (having or not the intention of being offensive) and the role of the harm caused on the audience with a claim. Since offensive meaning is a sort of evaluative meaning, our results can be at least partially extrapolated to the evaluative in general.

One of our results was that all factors –membership, status, intention and harm– have a significant effect in offensive meaning. Being out-group, having a high status, having a negative intention and causing harm on the audience make our statements more offensive than when we are in-group, low status, have a neutral intention and no harm is caused on the audience. This result is in line with the idea that meaning is generally tied to uses of language rather than to particular sentences.

Crucially, the speaker identity played a significant role in our studies, which in turn varied from vignette to vignette. For example, when the vignette was about a minority identity or feminism, the speaker identity effect was larger than in vignettes about other topics. This second outcome speaks in favor of two ideas. First, this result is in line with what social epistemologists and others have been recently emphasizing: we are socially situated subjects, and that fact has an important effect on the verbal and nonverbal actions we are able to perform (Ayala 2016, 2018; Kukla 2014). Second, the relevance of some topics, and our perception of them, are highly sensitive to the features of the particular society. Thus, both the relevance of a topic and other features of the context are decisive in identifying the meaning that someone communicates through her words. If we want to measure polarization in a society, how salient the topic is in that society must be taken into account, as well as other particularities of the context, in order to know exactly what kind of information is communicated by the participants of the tests used to measure polarization. In some cases, participants might express information about what they really believe, which contradicts what they sincerely say they believe. Other times, participants might express information not about what they really believe, but about what can be reasonably expected from them, i.e., their attitudes.

A third result of our studies was that participants showed a disanalogy in their responses when they were made in the abstract vs. when they were responses to particular situations. On the one hand, they ranked as more offensive the utterance of the same sentence when it was performed by an out-group speaker than by an in-group one, and this effect was larger than the effect of speaker intent, at least in our first study. So, speaker membership was taken by participants' judgments as a more important determinant of offensive meaning than intention. However, when they were explicitly asked about the factors' relevance in considering a particular utterance as offensive, they ranked speaker intent as the most relevant factor, while speaker membership was seen as the least.

This result can be seen as an instance of the idea that through our judgments we express the rule we actually follow, and the rule we actually follow may not be in accordance with the rule we sincerely say we follow. Participants said that they believe that the speaker intention is the most relevant thing in order to know whether their words were offensive or not. However, when they were asked to judge specific cases, they showed through their responses that they didn't think that the speaker intention is the most relevant factor. Our judgments express the rule we follow. And when they are evaluative in nature, they express our practical attitudes.

It is true that this idea contradicts with the one that generalizations about groups of people tend to be evaluations. One possibility is to say that generalizations about groups are instances of an exception. Other possibility is to say that claims about groups of people are not really generalizations, but claims about every member of the group. This second option would force us either to adopt an objectual interpretation of quantifiers, or to say that there is no quantifier involved in these claims. In any case, the idea that we tend to express, more often, the rule that we actually follow through our judgments in specific situations rather than through our judgments in abstract terms is just an advice: in order to try to measure people's attitudes at an early stage of polarization it seems better to provide enough context before asking them.

## 7.5. Conclusion

Along this chapter we have introduced the intuitive difference between the descriptive and the evaluative, a difference about uses of language rather than groups of terms, and have explained how it can be understood in relation to the Wittgensteinian approach to mental ascriptions introduced in chapter 6. Through the evaluative use of language we express our mind, our practical attitudes. Moreover, we have argued that our judgments in specific situations express more frequently the rule we actually follow than our claims about the rules we think we follow, and have shown that the evaluative meaning expressed through a claim is highly context-dependent. In this sense, we can express our attitudes even through a claim that appears to be descriptive a first sight, or through a belief self-ascription.

In chapter 6 we saw that our mind, our dispositional mental states, are conceptually linked to action. For instance, to believe that p is to have those commitments, conceptually articulated and linked with certain courses of action, related to p, which are determined by our social and linguistic practices, by our form of life. So, there is possibility of error regarding our mental ascriptions: it can be the case that the rule we say we follow is not the rule we actually follow. Through the evaluative use of language people express information about their attitudes, their practical commitments, given the action-guiding nature of the evaluative. In this sense, the distinction between the descriptive and the evaluative enables us to measure not what people say about themselves, but what they express. Thus, we can measure not only what people actually believe (which seems crucial to accurately measure ideological polarization), but also their affective attitudes, i.e., those practical attitudes closely connected to having a certain level of radicalism. In the next chapter, we will explain exactly how the philosophical assumptions introduced in chapter 6 and in this chapter allow us to reassess the concept of affective polarization in the way we do in this dissertation. More specifically, we will introduce our notion of polarization in attitudes and explain how it meets the desiderata for a suitable notion of polarization.

# Chapter 8

# Conclusion: Affective Polarization Reassessed

On July 7, 2020, *Harper's Magazine* published a letter signed by over a hundred and fifty famous journalists, authors and writers. The letter, entitled "A Letter on Justice and Open Debate" and known since its publication as the Harper's letter, aims to denounce the alleged censorship and restriction of freedom of expression suffered by most contemporary democracies. This letter found its equivalent in a Spanish letter in which the signatories adhere to the complaint, entitled "Carta española de apoyo al manifiesto Harper's". Our democracies, the signatories claim, are falling victim of the emergence of "a new set of moral attitudes and political commitments that tend to weaken our norms of open debate and toleration of differences in favor of ideological conformity". According to them, censorship, promoted by these attitudes "in response to perceived transgressions of speech and thought", is widespread, and it is constricting the free exchange of information and ideas. At first sight, advocating freedom of expression, one of the fundamental rights of any worthwhile democracy, is an issue that one might expect that people come around together rather than getting more divided over. However, this is not what has happened.

Against the supposed "cancel culture" denounced by the Harper's letter's authors, someone might have the impression that the authors endorse an epistemic libertarianism according to which any public contribution should be equally assessed if knowledge is going to be properly acquired (see Pinedo & Villanueva forthcoming, for a recent discussion of the idea of epistemic libertarianism). In the words of Harper's letter's authors, "The way to defeat bad ideas is by exposure, argument, and persuasion, not by trying to silence or wish them away". One might think that this principle is, say, a wolf in sheep's clothing. Given the inequality in social privileges enjoyed by certain groups, epistemic libertarianism presumably runs counter the initial objective of the right of freedom of expression, which was to give a voice to the voiceless. This is at least what many people have criticized of Harper's letter.

On July 10 of the same year, the online site *The Objective* published a letter in response to Harper's letter, called "A More Specific Letter on Justice and Open Debate", also signed by a group of journalists, authors and writers. In the response letter, the authors highlighted the misguided nature of Harper's letter, to put it mildly. Specifically, they emphasized that the signatories of the first letter were mostly white, wealthy and privileged people who were using one of the most prestigious and powerful magazines to complain that they are being silenced, which is practically a contradiction in itself. Secondly, the authors pointed out that Harper's letter missed the point in denouncing a situation that unfortunately is not new, and that people from certain social groups with low social power had been suffering since long ago. "The content of the letter also does not deal with the problem of power: who has it and who does not". Precisely, the authors observed that some of the signatories of Harper's letter had, in the past, unfairly harmed and criticized people from socially disadvantaged groups, thereby contributing to silencing certain individuals. In their words:

> The signatories, many of them white, wealthy, and endowed with massive platforms, argue that they are afraid of being silenced, that so-called cancel culture is out of control, and that they fear for their jobs and free exchange of ideas, even as they speak from one of the most prestigious magazines in the country. [...] Harper's is a prestigious institution, backed by money and influence. Harper's has de-

> cided to bestow its platform not to marginalized people but to people who already have large followings and plenty of opportunities to make their views heard. Ironically, these influential people then use that platform to complain that they're being silenced. [. . . ] The problem they are describing is for the most part a rare one for privileged writers, but it is constant for the voices that have been most often shut out of the room. When Black and brown writers are hired by prominent media institutes, NDAs and social media policies are used to prevent them from talking about toxic workplace experiences. [. . . ] We recognize a few of the signatories of the Harper's letter have been advocates of the issues that concern us here, which is, in part, the root of our hurt and dismay. Yet, everyone who signed the letter has reinforced the actions and beliefs of its most prominent signatories, some of whom have gone out of their way to harass trans writers or pedantically criticize Black writers.

So, one party vehemently complained that they were being silenced and persecuted, and did not listen to what the other side said; the other side claimed that complaining about being silenced when one belongs to a privileged group is in fact a way of perpetuating the injustices that some disenfranchised people suffer, and is not willing to consider the arguments provided by the other part. Nevertheless, some of the people now located in opposing sides concerning this issue have been in the same team in the past. Now, however, they have been split in a way which is similar, to some extent, to the situation that Anne Applebaum describes in her recent book. Many of those who were her closest friends in the late 1990s, a group of Polish conservatives with similar ideas, she says, are now separated from each other for political reasons: they shun each other, are unwilling to talk to each other, and would be embarrassed to acknowledge that they were friends in the past (Applebaum 2020: 9-23). Bridging the gap, something similar seems to have happened in the case at hand, suggested by the following statement made by those who write the letter responding to the first one: "We recognize a few of the signatories of the Harper's letter have been advocates of the issues that concern us here, which is, in part, the root of our hurt and dismay". Thus, this case can be considered as an instance of the kind of division Applebaum refers to in her book, which is dangerously widespread: "The estrangements are political, not personal. Poland is now one of the most polarized societies in Europe, and we have found ourselves on opposite sides of a profound divide, one that runs through not only what used to be the Polish right but also the old Hungarian right, the Spanish right, the French right, the Italian right, and, with some differences, the British right and the American right, too." (Applebaum 2020: 12). The case of Harper's letter seems to exemplify a situation of division between two groups, but the situation is not just a state of polarization, it is the result of a process, as in the Applebaum's case.

What leads a group of powerful and wealthy authors to compose and sign a letter strongly denouncing that they are being persecuted for expressing their opinions freely, and that consequently the boundaries of what can be said without the threat of reprisal are being narrowed? Does contemporary democracy find itself in a novel and dangerous situation of censorship, or is it rather a case of certain very privileged people displaying a negative attitude, as the other side claims? Are the authors of Harper's letter missing the point in claiming that they are silenced and persecuted, especially by not taking into consideration their social status and the evidence on the effects of offensive language in reinforcing certain unjust social norms? If so, why do they defend it so strongly and vehemently? Are they just doing that thing, missing a point, or something else? Why are both parts so confident in saying that the other side is wrong? Can this case be considered as an instance of polarization? In what sense?

In this chapter we will introduce the reassessed concept of affective polarization, that we will call *polarization in attitudes*. Polarization in attitudes has to do mainly with having certain attitudes closely linked to certain level of credence in the core beliefs of the group one identifies with. This notion of polarization is crucially tied to the philosophical assumptions introduced in chapters 6 and 7. In particular, it is tied to the following ideas. Regarding the assumptions resulting from chapter 6: There is often a gap between the rules one says one follows and the rules one actually follows, many beliefs are action-guiding, and there is contextual-authority regarding our own mental life. Regarding the assumptions resulting from chapter 7, those which are central here are the following: Evaluative uses of language are good indicators of our mind, the evaluative use of language extends beyond evaluative terms and is highly context-dependent, and our judgments on specific situations give more information about our mindset than our statements about the rule we follow.

If you were convinced by the notion of polarization in attitudes when we proposed it at the end of chapter 3, you have to know that you have committed yourself to the philosophical assumptions introduced in chapters 6 and 7. If these assumptions are not to your liking and you prefer to get rid of them, then you have the problems introduced in chapter 3, 4 and 5, and furthermore you cannot satisfy the desiderata we have proposed for a suitable concept of polarization in chapter 3. Along this chapter we will try to flesh out these ideas a little further, sometimes with the help of the case introduced at the beginning.

The general structure of this chapter is as follows. First, we characterize the notion of polarization in attitudes with the help of the tools we have obtained from the discussion of chapters 6 and 7. Second, we differentiate our notion of polarization in attitudes from other similar notions. Finally, we discuss the practical scope of our notion, its application. Here is what happens in this chapter in a bit more detail. In section 8.1, we characterize our notion of polarization in attitudes, make explicit the theoretical tools it needs from our discussion in chapters 6 and 7, and explain how it involves a reassessment of the concept of affective polarization. Section 8.2 discusses how our notion of polarization in attitudes meets the desiderata proposed in section 3.5. After this, we briefly discuss, in section 8.3, some of the attitudes related to the increase of polarization according to the literature on epistemic vices that seems similar to the attitudes that are central to our notion of polarization. More specifically, we discuss the attitudes of arrogance, dogmatism and closed-mindedness (section 8.3.1), and discuss whether they are necessarily bad attitudes and whether we are responsible for them (section 8.3.2). In section 8.4, we review some recent studies of affective polarization and analyze them from our notion of polarization in attitudes. Finally, in section 8.5, we briefly devise a possible design to measure polarization in attitudes, based on the recommendations that follow from our discussion along this chapter.

## 8.1. Reassessed, expanded and contextual affective polarization: Polarization in attitudes

The case introduced at the beginning of this chapter can be seen as an instance of a situation where two groups of people have opposing views and display certain attitudes, such as co-authoring a published letter to vehemently denounce the current situation, from one perspective or another. In that sense, this case can be seen as an instance of polarization in attitudes. Let us characterize the notion.

> **Polarization in attitudes**: Two people get more polarized in attitudes to the extent that, at a particular time in a specific society, they express to have certain attitudes, those with an especial link to action, that are closely connected with an increase in their level of confidence in the core beliefs of the group they identify with, which makes them more impervious to the reasons coming from the other side. Recall that the especial link to action exhibited by affective attitudes lies in the fact that one might not have the practical attitude that one expresses to have *only* if one is deemed irrational or incompetent. Otherwise, the speaker necessarily has the practical attitude expressed. On the contrary, one can be rational and competent and, at the same time, it can be the case that one does not have the attitude not specially linked to action that one expresses to have.

So, it can be argued that, in this case, each party self-identifies with a particular group of people (e.g., the group of people that is unjustly silenced; the group of people fighting against injustice) and has a high level of confidence in their core beliefs, which is showed through the things they do and the way they say what they say, i.e., their attitudes. The point is not simply that they say they believe one thing and they might actually not believe such a thing, but that through their statements, through their evaluative judgments, they express the kind of attitudes that have an especial link to action and that are connected with their level of radicalism. This is not only “missing a point”, as the authors of the second letter claim, but something else: it is to show how attached one is to a group. This diagnosis also applies to other cases discussed along this dissertation, such as the group of people that decided to demonstrate against the alleged existence of a criminal organization dedicated to falsely accusing teachers for financial gain in Ceuta (chapter 1), the group of people that decided to demonstrate with a golf club and a posh saucepan yelling that the Spanish government is responsible for the deaths caused by the COVID19 pandemic (chapter 2), and the group of people that dropped all kinds of hostile messages toward Carmena and her government for the changes made in the annual parade called “the Three wise men” (chapter 3). It is the increased confidence in the core beliefs of certain ideology that leads someone to publicly display such attitudes, which, in turn, are an indicator of such a level of radicalism. The attitudes that the Harper’s letter’s authors express by feeling the need to compound a letter to denounce the alleged oppression that people from their status suffer, as a means for fighting against the alleged constriction of the free exchange of information and ideas, are closely connected with their level of confidence in such ideas. To the extent that these ideas constitute some of the core beliefs of a given political ideology, it can be argued that through their actions these individuals may simply be expressing their strong adherence to those political ideologies.

This concept of polarization is crucially dependent on the theoretical tools we have introduced in chapters 6 and 7. The main tools of chapter 6 supporting this notion are the following.

> **Gap**: There is often a gap between the beliefs and other mental states that someone sincerely self-ascribes (the rules one says one follows) and the beliefs and other mental states in which one actually is (the rules one actually follows).

**Action-guiding mental states**: Beliefs, as well as many other dispositional mental states, are a set of commitments conceptually linked to certain courses of action. A subset of these attitudes exhibit an *especial* link to action, which can be called affective attitudes.

**Contextual authority**: Depending on the context, there are cases in which there is a presumption of authority regarding a mental self-ascription, cases in which the speaker exhibits a strong authority, and cases where there is neither strong nor presumptive authority.

The main tools of chapter 7 on which this notion is based are the following.

**Evaluative language as an indicator**: Evaluative language, in contrast to descriptive language, accomplishes the function of expressing the speaker's attitudes, those especially linked to action.

**Beyond evaluative terms**: The evaluative use of language extends beyond evaluative expressions and terms and is highly context-dependent, and then the pieces of speech that express information about the speaker's mindset, about her attitudes, also extends beyond a group of particular terms and expressions.

**The judging and thinking distinction**: Our judgments in specific situations are commonly more closely linked to our mind than our statements about what we think it is the rule we follow or the principle we embrace.

Having introduced all these theoretical assumptions, we are now in a position to see why the concept of affective polarization can be reassessed in the way we do here, as polarization in attitudes. Polarization in attitudes have to do with the kind of attitudes that have an especial link to action, those for which we frequently do not exhibit authority and that we express through our evaluative judgments. These attitudes, the information we express through the evaluative use of language, are calculable in virtue of our linguistic practices and the norms governing our social practices in general and are highly context dependent. Besides, these attitudes exhibit a peculiar feature regarding the possibility of error. The affective attitudes that someone says she has might not be those attitudes that she actually has, as in the case of attitudes that do not have an especial link to action. But the affective attitudes *expressed* by someone are conceptually incompatible with a later discovery that she does not have the attitudes she expressed, while this discovery is possible in the case of an attitude that does not have an especial link to action, as we have seen in the two previous chapters. Let us discuss our reassessment of the notion of affective polarization.

Recall the tools usually employed to measure affective polarization, introduced in section 3.2. We have distinguished between three tools that *directly* ask respondents to rate their feelings or make certain evaluations, and two other more *implicit* tools that measure respondents' behavior. The first three tools are the feelings thermometer, stereotype tests and feeling-linked-to-situation questionnaires. The implicit tools, on the other hand, are basically implicit bias tests and behavioral measures.

These direct tools, in contrast to the tools used to measure ideological polarization, always involve language that is commonly used to evaluate. The feeling thermometer asks people to rank how they feel in terms of warm (approval) and cold (disapproval) feelings regarding a particular issue, person or whatever. Feeling-linked-to-situation questionnaires also ask people to indicate how they feel in terms of positive and negative feelings. However, instead of asking about general topics or certain people without providing context, it asks respondents to indicate how they would feel in a very specific situation (e.g., if their daughter got married with someone self-identified as conservative/liberal). Finally, stereotype tests ask people to rate supporters of opposing political parties by ascribing them some adjectives from a particular set, which comprises positive and negative evaluative ones (e.g., "intelligent people", "betrayers of freedom").

It is our contention that, since these tools involve evaluative language, they usually measure what people express about their own mind rather than what they say of themselves. In other words, by self-ascribing a particular feeling toward a specific issue or by ascribing certain predicates to “the others”, respondents are often expressing their own perspective toward the topic or their opponents, i.e., their commitments especially tied to action, rather than simply reporting their own feelings or how their opponents are. Hence, affective polarization’s tools measure participants’ commitments, those conceptually tied to action they express to have, mainly in virtue of the theoretical tools *Gap* and *Evaluative language as an indicator*.

So, drawing on the *Evaluative language as an indicator* theoretical tool, we are able to assert that through the evaluative use of language we express our picture of the world, our attitudes especially linked to certain courses of action, what can be reasonably expected from us. In particular, and relying on the theoretical tool *Action-guiding mental states*, we can say that we express certain attitudes tied to a certain way of living that can be conceptually linked to having a high level of confidence in some core beliefs of certain ideological identities. In this sense, both the direct and implicit tools employed to measure affective polarization seem to measure, to a greater or lesser extent, what people express, people’s level of radicalism, namely, how impervious they are to the reasons and arguments provided by “the others”. Thus, a conclusion we draw from this is that we have to expand the range of evaluative language involved in these tools to measure polarization.[1] As we have said, someone’s linguistic and nonlinguistic behavior gives us information about their placement amongst social categories (see Davies 2020), especially through the evaluative use of language. For instance, claiming “When they call you ‘fascist’ that means that you are on the good side of history”, as the Spanish politician Isabel Díaz Ayuso recently claimed in an interview (Martiarena 2021), gives us a lot of information regarding what can be expected from the speaker, her attitudes, due to the social category, the socio-normative position, in which

[1] Interestingly, the results of a recent study conducted in Spain show that polarized people tend to speak about evaluative issues (e.g., political preferences and values) as if they were factual ones (Viciana et al. 2019). Once one has a very high level of confidence in one’s beliefs, one tends to see them as trivial assumptions, as facts. So, this could be used as a mark or sign of a high level of polarization. However, presumably this will not be very useful to measure certain polarization processes in their initial stage.

we can place her. And this information is not only about what the speaker actually believes, but about what can be expected from her, her attitudes. In contrast, making a descriptive claim such as "They called me 'fascist'" does not necessarily give us much information about the speaker's socio-normative position.

It is important to note however that, for example, saying that nowadays political correctness is killing free speech, or that climate change is a hoax, may also count as an evaluation, a claim that expresses much about the speaker's mind. But these claims don't contain any evaluative term or expression. The theoretical tool *Beyond evaluative terms* enables us to explain why: the attitudes we usually express through the evaluative use of language go beyond the use of certain terms and expressions. As we have seen in the previous chapter, evaluative meaning is highly context-dependent. So, the tools employed to measure reassessed affective polarization, polarization in attitudes, should not only be expanded to include other evaluative expressions than those relative to feelings, but also some evaluative uses of language in general. As we have seen in previous chapters, sometimes even by making a belief self-ascription one might make an evaluation and therefore express certain affective attitudes. That is what allows to explain why sometimes people can express their attitudes through the tools usually employed to measure ideological polarization.

According to *Gap*, one could, for example, sincerely say that one is in favor of gender equality, inclusivity, freedom of speech, etc., or that one believes that it is good to help people, and that none of this is the case. And according to the theoretical tool *The judging and thinking distinction*, it might be the case that, although in thinking about oneself one sincerely says that one is in favor of gender equality, inclusivity, etc., through one's evaluations in some specific situations one shows that one is not really in favor of gender equality or inclusivity, or that one is not a freedom of speech supporter but just a privileged person trying to conserve his privileges, or that one does not believe that it is good to help people. But, crucially, not only that: regardless of whether one's evaluations show that one does not believe what one sincerely says one believes, one's evaluations can simply express their affective attitudes, i.e., information about what can be expected from her from a practical point of view. Therefore, it seems to be a good strategy to grasp people's practical commitments to measure not only what people expresses, rather than what they self-report, but also to measure this in very specific situations. As Wittgenstein puts it, "my judgements themselves characterize the way I judge, characterize the nature of judging" (Wittgenstein OC § 149). In this sense, the feeling-linked-to-situation questionnaire, for instance, seems to be a better tool than the feeling thermometer to measure this type of polarization, because it provides more specific contexts for participants to make their evaluative judgments.

What is the particular perspective expressed by participants through the evaluative use of language, through their judgments? The right answer is that it depends, in part due to the theoretical tool *Contextual authority*, that is, the idea that there are contexts where there is a presumption of authority regarding a mental self-ascription, contexts in which the speaker exhibits a strong authority, and contexts in which there is neither strong nor presumptive authority . It depends on different factors of the context, such as who is the respondent, how the question is formulated, etc. But, in particular, it depends on how salient the topic in that society at that particular time is, and how central the issue is to certain political identities. The popularity of the discussion topic has a significant effect on polarization. A similar idea can be found in Burnstein and Vinokur (1977). The same topic can be perceived differently in different periods of time and can be more or less related to certain ideological identities. For instance, someone's self-report on the cold feelings she would feel if her daughter got married with a Muslim may indicate different things in different times and societies. In Spain, reporting having very cold feelings in that scenario may express a high degree of adhesion to certain ideological identities, and also a high level of credence in some of those ideological identities' core beliefs (e.g., the belief that Muslim people are taking advantage of Spain's resources, or whatever belief in this line maintained by far-right ideologies), to the extent that this issue has been a very salient one to some ideological identities. Of course, it could also simply express racism and not a particular level of adhesion to certain ideological identities. That is the main reason why what is expressed depends on the relevance of the topic in a specific society at a particular time, as we have said in relation to the example presented in the introduction of this dissertation (chapter 1). Besides, the theoretical tool *Contextual authority* enables us to explain why sometimes the tools employed to measure ideological polarization, and the feeling thermometer tool, can serve to measure this type of polarization.

Now, we would like to bring your attention to the different and possible types of affective polarization that we have distinguished in section 3.2.1. We have essentially distinguished four types: affective polarization via sympathy (APS), affective polarization with animosity (APA), affective polarization with radicalism (APR), and affective polarization with animosity and radicalism (APAR). Arguably, APS is the type of affective polarization that puts democracy less in danger, if it poses any risk at all. APA is possibly more dangerous than APS, but still, it is not clear how it can promote the problems that our contemporary democracies face. The reason is that, even though someone might think that supporters of Partido Popular political party are spoiled childish people who think that the world belongs to them, she might be willing to engage with their reasons and arguments, and reach consensus when needed. In this sense, this type of polarization does not necessarily endanger contemporary democracies. APR and APAR, on the other hand, seem to be types of polarization that potentially put democracy at risk. The main reason is that they comprise radicalism, i.e., a high degree of confidence in the core beliefs of the political group one identifies with, which means having certain attitudes. When one has a high level of credence in some beliefs, one becomes impervious to the reasons and arguments against those beliefs, because they are seen as nonsense, or as clearly false at best. Note that we have said that these types of polarization *potentially* put democracy at risk. The reason is that not necessarily every instance of ignoring others' arguments because we have a high level of confidence in our beliefs counts as a pernicious move. We will say something more in this line in section 8.3.2.

Let us now briefly take into account some of the possible problems that the concept of affective polarization might face, reviewed in section 3.2.2. The first objection was that it is neither exactly clear what the feelings that are measured by the feeling thermometer are, nor the reasons behind indicating such feelings. The second one was that it is not clear how the different phenomena measured by the tools of affective polarization are connected, if they are.

With the reassessment of the concept of affective polarization that we propose, the feeling thermometer does not measure the feelings that people have in phenomenological terms, but, sometimes, it measures the attitudes participants express, which are ascribable to them in virtue of the rules governing our practices at that particular time. Certainly, what is measured by the feelings thermometer tool is highly context-dependent (see chapter 7). At times, what it measures is closer to what people say of themselves than to what they express –as it happens with the tools employed to measure ideological polarization. But it does not measure feelings in a phenomenological way. In this sense, the first objection seems avoidable. The second objection can also be avoided insofar as it can be argued that certain behavior, on the one hand, and the indication of certain feelings toward certain issues, on the other hand, are connected at a particular time (see chapter 6). If I say that I would feel very cold feelings if my daughter got married with a Muslim (regardless of my phenomenology), presumably I will behave in a certain way toward Muslim people, especially if it is a salient political issue. Our dispositional mental states are connected with certain courses of action, and our evaluations are especially connected with action, with certain attitudes.

The third objection was that, if the essence of affective polarization lies in the feelings we have toward certain people, then it seems that the rise of polarization is an irrational phenomenon, and this thesis, although held by some authors, does not seem compatible with other ideas, such as that people necessarily think that they have the truth by their side, and produce arguments to support their positions. Since this third reason has to do with one of the desiderata we have proposed for a suitable concept of political polarization, we will leave it to section 8.2. However, our answer has somehow already been introduced.

Finally, we want to end this section by pointing out what someone who rejects the philosophical assumptions we have introduced in chapters 6 and 7 would have to say regarding the concept of affective polarization. If one rejects these assumptions, one seems to be forced to maintain that (i) all tools employed to measure affective polarization would have to give the same result with the same sample of people, otherwise the respondents are irrational. The reason is that if one assumes first-person authority, the different tools used to measure affective polarization should obtain the same results when applied to the same sample of people or those people are irrational. (ii) Affective polarization is only about feelings and not about other attitudes closely connected with having certain degree of belief, which favors the irrational explanation of polarization. Our diagnosis, although focused on the affective attitudes that people express, is connected to the level of confidence in certain beliefs and with many other inferences that can be used to rationally explain why someone has certain attitudes. (iii) The tools employed to measure affective polarization should work for any society at any time, because they only measure what people say, that always coincide with what they actually feel. (iv) The different types of affective polarization we have introduced cannot be distinguished. (v) The notion of affective polarization still faces the problems and limitations, introduced earlier, associated with the notion, and therefore it cannot meet the desiderata we have proposed. In the next section, we explain how our notion of polarization in attitudes meets the proposed desiderata, and briefly discuss which of them cannot be satisfied without the philosophical assumptions of chapters 6 and 7.

## 8.2. Meeting the desiderata

The aim of this section is to explain how the concept of affective polarization reassessed in the way we propose, that is, polarization in attitudes, meets the desiderata put forward for a suitable concept of polarization, introduced in section 3.5, with the help of the theoretical tools provided by chapters 6 and 7. Moreover, we also discuss which of those desiderata cannot be satisfied if the philosophical assumptions introduced in chapters 6 and 7 are rejected.

Let us start with the first desideratum proposed, which was the one we have called EVIDENCE: A suitable notion of polarization must be consistent with the best available evidence. As we have seen, some types of affective polarization are characterized by people's attitudes, in particular by people's disposition to disregard others' reasons because of their high level of credence in the core beliefs of the group they identify with. It is people's behavior, verbal and nonverbal, in specific situations, and not their mental self-ascriptions, what allows us to know more accurately their level of confidence in certain beliefs. These types of affective polarization, or polarization in attitudes, involve radicalism, which can be yielded by different mechanisms, such as identity psychological mechanisms, group membership, party sorting, etc. But the mechanisms promoting radicalism do not only appear within groups of likeminded people and mustn't be addressed separately from the other evidence available (see chapter 4). Being exposed to others' arguments may get us more divided under certain conditions, especially when we are able to filter the information we are exposed to, leaded by our motivated reasoning; but also because of the configuration of the system we inhabit (Dorst 2020) and because of certain kinds of public discussion, in particular those that count as crossed disagreements. These mechanisms affect the pool of arguments we are exposed to, in particular its size and density, which, as a consequence, increase our degree of belief in our previous positions. Being mainly exposed to the repetition of a set of arguments that confirms what we already believed makes us more confident in our views. That is radicalism. According to our review of the main evidence concerning how we polarize, radicalism seems to be better positioned than extremism in order to accommodate it (see chapter 4). But in order to argue that affective polarization is about the level of radicalism we display through our attitudes and not about our mental self-reports, we need at least *Gap*, *Action-guiding mental states*, *Evaluative language as an indicator* and *The judging and thinking distinction*. The idea that affective polarization, polarization in attitudes, involves radicalism can only be held if the following assumptions are in place: there is a gap between the beliefs we sincerely self-report and the beliefs in which we actually are, many of our beliefs are closely connected with our verbal and nonverbal behavior, through the evaluative use of language we express our attitudes especially tied to action, and our evaluative judgments in specific situations often express the rule we actually follow more accurately. In this sense, polarization in attitudes seems to meet this desideratum. If these and the other philosophical tools supporting the reassessment of the concept of affective polarization are rejected, then it must be held that affective polarization just has to do with having certain feelings, and that diagnosis can hardly accommodate the available evidence on how we get polarized, which mostly has to do with the increase in confidence in certain beliefs. Moreover, the semantic theory introduced in chapter 7 enables us to accommodate another piece of evidence that is crucial for our purposes, the intuitive distinction between the descriptive and the evaluative. Without this theory, or another one able to accommodate the requirements introduced in chapters 6 and 7, this desideratum can hardly be met. Expressivism is just an example of the fact that there are semantic theories that accommodate these requirements.

A second desideratum proposed was DANGEROUSNESS: A suitable notion of polarization must be consistent with the pernicious effects for democracy of a high level of polarization. As we have seen, at the mass level, polarization leads people to regard the arguments and reasons of their political opponents as misguided and as a threat, and to evaluate those perceived as "the others", as dishonest, unintelligent, etc. (Talisse 2019: 95). Polarization also increases distrust in public institutions and in government, increases intolerance, and corrodes the proper functioning of democratic institutions (Carothers & O'Donohue 2019: 1-2). According to Sperber and other scholars, a reliable informant must meet two conditions: she must be competent and benevolent (Sperber et al. 2010: 369). However, how competent and benevolent an informant is depends partially on how competent and benevolent she is perceived to be. If our confidence in a particular belief is too high, we tend to perceive people that think otherwise as incompetent, because our level of confidence in that belief leads us to think that, since the truth of the content of our belief is so evident to us, those who do not believe it must be incompetent. Another option is to think that they are competent but malevolent; they know that what we believe is true but have some perverse intention and that is the reason why they try to convince us that we are wrong. In both cases, we will not take as reliable informants those who have a different opinion from ours, getting more credibility to what is said by ingroup people, and increasing our adhesion to our group (Ortoleva & Snowberg 2015), i.e., increasing our confidence in the core beliefs of our group. That explains why polarization in attitudes implies being impervious to others' reasons and potentially generates the negative consequences noted above. Again, only with the help of the theoretical tools we have introduced in chapters 6 and 7 we are able to reassess the concept of affective polarization in terms of radicalism (see section 8.1). Certainly, these pernicious effects for democracy can be explained even if we reject the philosophical tools we have adopted. In terms of feelings and hooliganism one might explain why coordination is so hard to reach and why there is much hostility in a highly polarized society. However, without our philosophical tools, it is hard to explain why some contemporary democracies exhibit these problems and, at the same time, according to the results obtained by employing tools such as the feeling thermometer tool, the same society is allegedly not polarized. The attitudes that put in danger democracy can increase and at the same time the population of such a democratic society might respond to the thermometer tool in a way that this increase in attitudes were imperceptible. In this sense, our notion is better positioned than the traditional understanding of affective polarization to meet this desideratum.

Another proposed condition that a suitable concept of polarization must meet was RATIONALITY: A suitable notion of polarization must neither blame people nor account for the issue in terms of irrationality. The reason to propose this condition was that a rational story of polarization processes seems more compatible with the big picture of the main available evidence of how we get polarized than an irrational one (Dorst 2020). But this desideratum has also a political flavor. As we have said in sections 3.5, the pursued concept of polarization should not explain the increased polarization in terms of irrationality or lack of interest in truth. This line of explanation blames people at worst, and is incomplete, and even wrong (Dorst 2020)), at best. On the contrary, a suitable concept of polarization must be able to explain why getting polarized is a rational process given the current situation, acknowledging that it is precisely the fact that people care about truth, and that people give reasons supporting their own beliefs, that lead them to polarize; polarization is not just a matter of irrationally disliking the other part. Of course, this does not mean that we citizens have no responsibility for the current condition of our democracies, or that we cannot do anything. As we will see, we can be responsible in different senses (Cassam 2019; see section 8.3.2). But this point is compatible with the story according to which polarization is not an irrational, but a rational process: people's thinking is not necessarily riddled with irrationality.

We have argued against the sharp distinction between belief-like and desire-like mental states regarding their connection with action (see chapter 6), and maintained that affective polarization, or polarization in attitudes, indeed has to do with the attitudes linked to a certain level of confidence in one's ideological identity core beliefs. In this sense, being polarized is not a matter of irrationality. Our verbal and nonverbal behavior is closely connected with our mind, with our beliefs and attitudes, with the way we believe what we believe (*Action-guiding beliefs*), and with the attitudes linked to certain confidence in certain beliefs expressed through our judgments and behavior. In this sense, our attitudes are not just the result of our feelings toward the in-group and the out-group, but the result of our level of confidence in our beliefs, which explains why it is rational to disregard the reasons coming from the other side when the level of radicalism is high. Besides, if we take together the findings from psychology, political science and philosophy, instead of focusing only on the psychological mechanisms underlying polarization processes, it can be argued that the irrational picture of polarization –according to which we are mainly biased, dogmatic and arrogant people– is not so compelling (Dorst 2020). On the contrary, all the available evidence, taken together, suggest that, given the current environment of information flow, crossed disagreements and other linguistic phenomena, polarization is the result of being a rational person. Polarization can be seen as the result of being a rational person, rather than the opposite. It is rational to trust those that are on your own side, especially when you have a high level of confidence in the core beliefs of your group. And it is rational that, given the dynamics of information consumption, you end up reinforcing your own initial positions (see chapter 4).

The fourth desideratum proposed was DISANALOGY: A suitable notion of polarization must accommodate the disanalogy between self-ascribing a mental state and expressing a mental state in order to accurately measure it. An operational notion of polarization must be able to accurately measure people's mental states, more specifically those related to the bad consequences of the rise of polarization. As we have suggested, people's sincerity is not enough to guarantee that they are in the mental state they say to be in, especially when talking about complex issues and in abstract terms. In that sense, the first-person authority thesis seems challengeable, and therefore the concept of polarization must accommodate the disanalogy between self-ascribing a mental state and expressing a mental state and should measure the level of adhesion to a particular ideological identity. Our Wittgensteinian approach to mental ascriptions enables us to accommodate two possible types of error when self-ascribing a mental state: we might not be in the mental state we say we are in, but we might also express something else by self-ascribing such a mental state, that is, our affective attitudes. In the previous chapters, we have argued that the tools commonly used to measure affective polarization actually measure the attitudes that people express rather than just their emotional states. Our concept of polarization in attitudes endorses the approach to mental state ascriptions introduced in chapter 6 and the approach to certain uses of language discussed in chapter 7 (see section 8.1). In this sense, it is quite obvious how our notion meets this desideratum: the theoretical tool *Gap* is precisely the assumption that there might be a difference between our mental self-ascriptions and the mental states in which we actually are as well as the attitudes we express through our statements. Those who reject this assumption and assume the thesis of first-person authority cannot satisfy this desideratum.

The last condition proposed was INTERVENTION: A suitable notion of polarization should allow us to develop mechanisms to intervene as soon as possible. We have tried to clarify what is really measured by the tools commonly used to measure affective polarization, what types of affective polarization can be distinguished, which of them may imply certain bad consequences for democracy, and we have offered some recommendations that allow us to detect the type of polarization that endangers our democracies as soon as possible. Thus, our reassessed concept of affective polarization meets this desideratum to the extent that we have been successful in doing all this. If, as a result of our discussion, we are now better positioned to measure the dangerous type of polarization in a more accurate way, then it can be stated that our contribution was a kind of intervention, because it enables us to measure polarization more accurately, and then measure it as soon as possible. Moreover, reaching a deeper comprehension of the type of polarization that endangers democracy also enables us to devise new intervention strategies specifically aimed at ameliorating particular situations with certain specific features. But the crucial point here is that our recommendations to measure polarization in a more indirect way, by attending to what people express and not to their mental self-reports, enables us to measure polarization before the level of polarization is too high. This is crucial because the higher the level of polarization, the lower our chances to depolarize. According to our notion, the feeling thermometer tool, for instance, is useful only when the level of polarization is very high (and something similar occurs with the tools employed to measure ideological polarization) and in that sense it leaves us little room for intervention (see section 8.4). So, our reassessed notion enables us to intervene because it enables us to evaluate which of the available tools are better than others to measure polarization as soon as possible, and to design new ways of measuring polarization to detect it when the curve of polarization is not yet too high and then our possibilities to depolarize are bigger. The attitudes people express, those especially linked to what can be expected from them and that are connected to their level of radicalism, cannot be identified nor measured if one does not assume the philosophical tools introduced in chapters 6 and 7. Our notion needs to assume the possibility of error when talking about our mind as well as to accommodate the distinction between the descriptive and the evaluative. More specifically, these assumptions are needed in order to argue that our affective attitudes, those connected to our level of radicalism and more frequently expressed through our evaluative judgments in specific situations rather than through our direct mental self-ascriptions, are what should be measured to account for certain types of pernicious polarization as soon as possible. In this sense, our notion satisfies this desideratum.

After characterizing our notion of polarization in attitudes and explaining why it can meet, with the help of the theoretical tools introduced in chapters 6 and 7, the proposed desiderata, in the next section we will discuss different notions of attitudes that, although similar, are different from our notion, and also discuss whether we are always responsible for having such attitudes.

## 8.3. Epistemic vices: Attitudes related to the rise of polarization

In this section, we discuss certain attitudes that are very similar to the attitudes we point to with our notion of polarization in attitudes, but that are not exactly the same ones. These attitudes are those of arrogance, closed-mindedness and dogmatism, much discussed within the literature on epistemic vices. Commonly, these attitudes are associated with irrationality.

In the field of virtue epistemology it is common to examine agents' virtues, and more recently also vices, regarding the acquisition, retention and transmission of knowledge. Epistemic virtues are those that favor the acquisition, retention or transmission of knowledge, while epistemic vices are those that systematically get in the way of knowledge (Medina 2013) at some of the three mentioned levels (Cassam 2019). Cassam has recently distinguished three types of things that can be deemed epistemic vices: character traits, ways of thinking, and attitudes (Cassam 2019: 12-13).[2] Character traits are stable dispositions to act, think and feel in certain ways. Ways of thinking are particular reasonings, instances of thinking in a particular way. Attitudes are orientations or postures toward something, similar to character traits but less stable in time. For instance, a person can display the epistemic vice of arrogance because she is an arrogant person, because she is thinking in an arrogant way, or because she exhibits arrogance with respect to certain aspects. These are three distinguishable forms of epistemic vices. Note that this notion of attitudes is narrower than the one we introduced in chapter 4 and to which we appeal when we speak of attitudes along this dissertation.

Recently, the topic of epistemic vices has received significant attention in relation to the rise of political polarization. In particular, it has been analyzed how some different epistemic vices might have been the causes of the current level of polarization in many democracies (see, for instance, Lynch 2019; Tanesini & Lynch 2021). Some of the most outstanding epistemic vices in this sense are arro-

[2] This conceptual distinction is reminiscent of that made by Gilbert Ryle regarding emotions (Ryle 2009). Ryle distinguished between character traits (e.g., being a sad person), moods (e.g., being sad) and feelings (e.g., feeling sadness).

gance, closed-mindedness and dogmatism. We devote this section to discussing these three epistemic vices as attitudes closely related to the rise of polarization in attitudes.

However, contrary to many of the proposals, we start from the assumption that these attitudes do not need to be conceived as causes of polarization nor as the keys pressing in favor of the irrational story of polarization. Maybe they are just some consequences of other mechanisms that indeed foster polarization, and in that sense are signs of the rise of polarization. As we have seen, crossed disagreements, our capacity to filter information, the sorting phenomenon, etc., can foster the confidence in some beliefs central to the ideological group we identify with. But, in certain situations, it can be rational to increase our confidence in those beliefs. In this sense, the attitudes associated with being arrogant, closed-minded or dogmatic may be about what polarization consists in rather than about the attitudes that bring about the rise of polarization. As we have previously stated, this second option seems better to us for two reasons: it seems more compatible with all the evidence taken together and avoids blaming people for the current condition of many societies. Note that the terms 'arrogance', 'closed-mindedness' and 'dogmatism' have an evaluative flavor. Of course, a particular attitude can be described using these labels in a way that blocks the evaluative meaning usually expressed through them. But still, putting too much emphasis on these epistemic vices when talking about polarization conveys the risk of ending blaming people. At the end of this section, we will say something about the type of responsibility that can be demanded from us as polarized citizens.

### 8.3.1. Closed mindedness, epistemic arrogance and dogmatism

One of the philosophers that have recently analyzed the role of arrogance concerning the rise of polarization is Michael Lynch, whose diagnosis points out that most of us behave like a know-it-all, and that leads us to polarize. Briefly, Lynch's diagnosis lies in that a tribal or group-indexed epistemic attitude (Lynch 2021: 141-154, Lynch 2019), namely intellectual arrogance, "is bound up with, and worsens the effects of, affective or attitude polarization" (Lynch 2021: 141).[3] According to Lynch, intellectual arrogance is an unwillingness to regard one's worldview as capable of improvement from the evidence and the arguments coming from the other side. The idea is that when one is an arrogant, one becomes impervious to the others' reasons. In particular, Lynch's analysis of intellectual arrogance has to do with putting ego before truth, being hyper concerned for one's self-esteem, and being fear of error and defensive (Lynch 2021: 143). Intellectual arrogance is delusional in nature, says Lynch, and that is the reason why people rarely see it in themselves. This attitude can become tribal, in the sense that it can be experienced as part of a "we" and directed at a "them". Moreover, this epistemic attitude of taking one's own beliefs, in particular the beliefs the group one identifies with, as epistemically unimprovable, is an irrational attitude, says Lynch (Lynch 2021: 146). This type of attitude explains the pernicious aspect of polarization: "If a social group A arrogantly regards itself as epistemically superior to some group B about some subject S, then they will regard B as less trustworthy, reliable, or informed about S" (Lynch 2021: 146). So, if tribal intellectual arrogance increases, polarization increases. To put it another way, intellectual arrogance plays a key role in deepening our disagreements both over policies and attitudes, and hence in a type of polarization: polarization in attitudes. Certainly, this attitude is one of those attitudes that seem to be closely connected with having a high level of confidence in the core beliefs of the group one identifies with. But, to point out just two possible differences with our notion: first, we do not think that becoming impervious to the other's reasons is necessarily an outcome of an irrational process, nor that it has necessarily to do with putting ego before truth. Second, one can behave in an arrogant way without noticing it and, therefore, it is not necessarily the result of a group regarding itself as epistemically superior to another group.

A second attitude related to the way we have understood the type of polarization that endangers democracy is the so-called "closed-mindedness". According to Fintl, someone is closed-minded when is unwilling to be persuaded by the ar-

[3] Tanessini characterizes this attitude as "an unwillingness to submit oneself to the norms governing ordinary conversation and rational debate" (Tanesini 2016: 85).

guments of others (Fantl 2018: 12; see Allen 2020 for a recent analysis of Fantl's analysis). In a similar vein, Cassam defines a closed-minded individual as one that is disposed to freeze on a given conception, to be reluctant to consider new information, and to be intolerant to those opinions that contradicts her own (Cassam 2019: 33). More recently, Battaly has defined closed-mindedness as "an unwillingness or inability to engage (seriously) with relevant intellectual options" (Battaly 2021). Note that this epistemic vice is closely related to epistemic arrogance. However, the main difference between them lies in the fact that someone can be unwilling to be persuaded by others' reasons and evidence without thinking that one's position is not subject to improvement and behaving as a know-it-all. Simply, one is not open to consider other possibilities.

It is noteworthy that the boundaries between the attitudes called arrogance, closed-mindedness and dogmatism are not so sharp. One way to see it is that arrogant and dogmatic attitudes are subsets of closed-mindedness. This is so because it seems that you cannot be arrogant or dogmatic without being closed-minded, but you can be closed-minded without being arrogant or dogmatic. If you are intellectually arrogant, then you are closed-minded because you are not willing to consider other options. But you can be closed-minded without being arrogant. Similarly, if you are dogmatic, then you are closed-minded too, but you can be closed-minded without being dogmatic. What is it to be dogmatic?

Dogmatism is the third attitude mentioned above that seems to be displayed by polarized people. Kidd conceives it as a "disposition to respond irrationally to attempts by others to offer instruction and criticism" (Kidd 2021: 63). Battaly emphasizes its capacity to hinder our willingness to take others' position as serious possibilities: "it is an unwillingness to engage (seriously) with relevant alternatives to a belief one already holds" (Battaly 2021). According to Cassam, "dogmatism is more limited in scope and pertains to one's doctrinal commitments rather than to one's epistemic conduct generally" (Cassam 2019: 109). These authors recognize that dogmatism is somehow a type of closed-mindedness in the sense that it involves a kind of indoctrination or strong identification with the core ideas of a particular identity or group. At any rate, the important thing for us here is just that these attitudes, usually referred to in the literature by the labels of 'arrogance', 'dogmatism' and 'closed-mindedness' seem to be closely related to attitude polarization.

Certainly, the attitudes of being unwilling to regard one's view as capable of improvement and to be unpersuaded by the arguments from the "other side" are closely related to the diagnosis we have tried to outline along this dissertation. It is a high level of credence in the core beliefs of our ideological group, namely radicalism, that can lead our democracies to collapse, because it makes us impervious to others' reasons. In this sense, it might be useful to try to measure the level of presence of these attitudes in a society in order to grasp the level of polarization of that society. However, the presence of certain level of these attitudes does not need to be understood as a situation that is the result of, and driven by, people's irrationality, disregard for the truth and selfish interests, as some authors suggest. Nevertheless, stating that getting polarized can be seen as a rational process does not necessarily mean that we have no responsibility for having the attitudes we have when we are highly polarized. In the next section, we briefly discuss whether being impervious to the other's arguments, in different ways, is always a bad behavior or, on the contrary, there are situations where there is no responsibility that can be demanded from us.

### 8.3.2. Are these attitudes necessarily bad ones?

Some authors have argued that remaining steadfast in one's beliefs has epistemic advantages in group deliberation (Hallson & Kappel 2020; Levy 2019, 2020). Others, on the contrary, have argued that these characteristics can lead groups to end up in stalemates and to polarize (see, for instance, Tanesini 2021), although they recognize that in certain situations remaining steadfast can have certain advantages, in particular when the disagreement is transient. Others have argued that the alleged epistemic advantages of these attitudes are not in fact consequences of such attitudes, but of very similar ones, such as firmness (Cassam 2019), and hence that these attitudes are epistemic vices that almost always get in the way of knowledge. In this sense, these attitudes should always be avoided (Cassam 2019).

We are more inclined to a fourth option here. Recall that, although being unwilling to revise our own beliefs in the face of alleged evidence against may be pernicious because it can lead to the problems associated with the rise of polarization, to engage with others' reasons can also get us more divided in certain situations, as we have seen in chapter 4. So, according to a fourth option, we are sometimes epistemically entitled to ignore the alleged evidence that contradicts our beliefs, and not only in transient situations. The reason for that is that not all opinions must always be equally taken into consideration. There are at least two routes that can be followed in order to flesh out this stance.

One is to consider that, although being open to revise our own beliefs is usually epistemically beneficial, there are topics regarding which it is not advisable to stand open-minded, specifically when those topics carry certain political implications. In Allen's words, "[w]hile we epistemically ought to be open-minded in general, the importance of being open-minded is roughly proportional to the moral, social, or political significance of the matter at hand" (Allen 2020: 3). So, this first option has to do with the contents being discussed, in particular with their political significance. For instance, if taking into consideration certain positions on abortion would carry some political implications that are unacceptable, then we can stay closed-minded regarding such a topic and in consequence ignore those positions. This option can be broadened by including not only topics with unacceptable political consequences, but also topics in which the acceptance of certain positions challenges many of our assumptions that are necessary for many of our well-established practices. For instance, accepting that the Earth is flat, or that ghosts, as immaterial things that can nevertheless affect material things, exist, implies having to change many of our basic physic laws and assumptions that are a necessary ground for a big amount of our practices. Then, in this case, according to this first option, we can also remain closed-minded.

The other route would be to argue that, given the huge amount of information we are exposed to and that our capacity to attend to it is limited, not every opinion has to be necessarily taken as worthy or as information we are forced to engage with: the right to be heard in public must be earned, and can be lost too. In this sense, we do not ought to be epistemically open-minded in general;

it will depend on each particular case, and not only because of the possible unacceptable consequences of taking into consideration certain positions. That is, each case has to be evaluated not only in terms of the contents at hand, but, for example, also in terms of who the speaker is. In Pinedo's and Villanueva's words, "When confronted with new evidence, it's not the case that everyone has an a priori right to turn their opinions into epistemic possibilities that cannot be properly ignored. Being able to take part in a meaningful epistemic discussion is a right that can be earned, and it can be lost as well" (Pinedo & Villanueva forthcoming: 14). It is important to note that this fourth option is especially concerned with offering epistemic policies that serve as forms of epistemic resistance and, in that sense, it presupposes that injustice must be fought against. Our society can be represented as an unjustly organized socio-normative space where each node is associated with a limited number of possibilities for action (Ayala 2018; Haslanger 2015). It is the recognition that not all nodes in the socio-economic space are equal, together with the need to combat injustice, that must be especially considered when assessing whether or not a particular public contribution is worthy of consideration.

To end this section, let us return to the attitudes related to radicalism, that is, those that have to do with a high degree of adherence to a political ideology that makes people have a very pernicious level of credibility in the core beliefs of their group. To the extent that polarization processes are understood as rational processes, it does not seem to make sense to blame people for being polarized. But do polarized people exempt from any kind of responsibility? Can't their attitudes be not only reprehensible, but even blameworthy in certain cases?

Cassam separates blameworthiness from reprehensibility concerning the responsibility people have over their epistemic vices: "blame is not the only form of criticism, and it is possible to be critical of a person's epistemic vices without blaming them. Whether or not a deeply arrogant person deserves blame for being that way, they can certainly be criticized for their arrogance" (Cassam 2019: 6). In addition, he distinguishes two sorts of responsibility: acquisition responsibility and revision responsibility (Cassam 2019: 18-20). The first one is a type of responsibility that is bound to the way we acquire or develop the epistemic vice. If our past deliberate decisions led us to acquire a particular epistemic vice, a particular attitude, then we are acquisition responsible for it. The second type of responsibility is bound not to the way we acquire an epistemic vice, but to our ability to modify it. Revision responsibility is dependent on whether epistemic vices are malleable enough for revision. That is, if we have certain type of control over them, then we are revision responsible for our epistemic vices. Thus, we have five conceptual possibilities. One might not be responsible at all for one's attitudes insofar as they are not the result of one's past decisions and cannot be modified. But when there is some kind of responsibility for one's attitudes, one might be acquisition blameworthy, revision blameworthy, acquisition reprehensible, and revision reprehensible.

Thus, even though our reassessed concept of affective polarization either avoids blaming people for the rise of polarization as the irrational story would have it, it is still compatible with saying that certain attitudes are blameworthy or reprehensible to the extent that the agents displaying those attitudes are considered to have one or another type of responsibility.

In this section we have discussed some of the attitudes that might be considered as the affective attitudes tied to having certain level of radicalism that characterizes our notion of polarization in attitudes. In the next two sections, we are going to focus on the third part of this chapter: the practical dimension of our notion of polarization. In particular, in the next section we will discuss what can be said, from our notion, about some recent studies of affective polarization. After that, in section 8.5, we will offer a possible sketch of some vignettes aimed at measuring polarization in attitudes, following our recommendations.

## 8.4. Discussing some recent polarization studies

In this section, we briefly discuss some recent studies of affective polarization based on our results so far. In particular, we review the design of three cross-national studies and three other studies focused on Spain's level of affective polarization. Most research on affective polarization, including the ones discussed here, analyzes the level of affective polarization in a society through the feeling thermometer tool. Moreover, the questions aimed at measuring affective polarization are presented to the survey's respondents along with many other questions designed with different purposes, which might be problematic. Let us introduce some of these recent studies and discuss them from our notion of polarization in attitudes.

Gidron, Adams and Horne have recently conducted a comparative study in which they analyze affective polarization levels across twenty Western societies over the past three decades (Gidron et al. 2020). The countries are Australia, Austria, Canada, Denmark, Finland, France, Germany, the United Kingdom, Greece, Iceland, Ireland, Israel, the Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland and the United States. To do so, they analyze over eighty national election surveys that include, among many other things, questions about respondents' feelings toward political parties of their country. These surveys, and their results, are obtained from the Comparative Study of Electoral Systems (CSES). The question corresponding to feeling thermometer reads as follow: "I'd like to know what you think about each of our political parties. After I read the name of a political party, please rate it on a scale from 0 to 10, where 0 means you strongly dislike that party and 10 means that you strongly like that party" (Gidron et al. 2020: 14-15). Specifically, they measure levels of affective polarization by paying attention to the average of out-party dislike and in-party liking.

Certainly, in virtue of the theoretical tool *Evaluative language as an indicator*, our notion enables us to say that indicating that one strongly likes or *dislikes* certain political party might be a way of expressing one's commitments to one's own political identity, one's attitudes especially linked to action, and, in that sense, it might be a measure of how impervious to the others' reasons respondents are. However, the feeling thermometer tool is the most direct one of those commonly used to measure affective polarization, because it directly asks people not to make a judgment in a specific situation, but to indicate, after thinking, what they feel about someone or about certain topic. In virtue of the theoretical tool *The judging vs. thinking distinction*, our notion warns us that the feeling thermometer tool is more about what people say than what people express, even though the tool involves language that is commonly used to evaluate. Thinking about my feelings toward a particular political party often requires reflecting on the rule that I think I follow, without any specific situation to rely on. Thus, in virtue of *Gap*, the attitudes I say I have and the attitudes I actually have could be different. Presumably, only when the level of affective polarization is already quite high, one is willing to say that one has very cold or very warm feelings toward something or someone, and then to express their attitudes, closely tied to their level of confidence in the core beliefs of the group they identify with, through their responses. Only in those cases the feeling thermometer tool might serve to measure the attitudes expressed by participants. Therefore, our notion diagnoses that this way of measuring affective polarization is useful only when the polarization curve is already very high. Besides, it is to be expected that what the participants express through their responses is highly context-sensitive; it is dependent on the particularities of the study, the society, the time, etc.

Another comparative study, conducted by Boxell, Gentzkow and Shapiro, has also measured trends in affective polarization in several countries over the past four decades (Boxell et al. 2020). In particular, they measured it in nine countries which are members of the Organization for Economic Co-operation and Development (OECD). These countries are Australia, the United Kingdom, Canada, Germany, New Zealand, Norway, Sweden, Switzerland and the United States. To do so, they have constructed a new database from 116 different surveys in which there is a reasonably continuous series of questions related to people's feelings toward other political parties (Boxell et al. 2020: 3-4). In this case, the questions about respondents' feelings vary across surveys, but they commonly ask about respondents' feelings toward a particular political party, i.e., feeling thermometer. Despite the fact that there are some interesting differences between this study and the study conducted by Gidron, Adams and Horne, we think that the same can be said here.

A less recent but more interesting (for the purposes of this dissertation) comparative study of affective polarization was conducted by Westwood and other scholars across four countries (Westwood et al. 2018). The countries were Belgium, the United Kingdom, Spain and the United States. The interesting thing about this study is that in this case, researchers did not employ the feeling thermometer tool, but a more indirect one. In particular, they designed a version of the classic trust game. In this game, participants are given a cash allocation and can give all, some or none of the money to another participant. The rules are that the amount of money given to the other participant will be tripled, and the second participant can return all, some or none of the money back to the first one. Thus, the more trust deposited in the other participant, the more should be allocated to her and the more can be gained back (Westwood et al. 2018: 340). In this particular version of the game, participants could see the profile of the other participant, which includes information about her party identification along with other information (participant's gender, age and income) aimed at making the experiment less obvious. As expected, the results of the experiment showed that partisan identification has a great impact on trust, even over other social factors like ethnicity or religion: participants gave more amount of money to participants that belong to the same political party or to another ideologically similar one, and less to those that self-identify with a political party of the "other side". An interesting finding of this study was that, contrary to the studies previously introduced, the United Kingdom is not as homogeneous a society as it is commonly stated: both social and political clues of the other player lead participants to trust them more or less. In fact, this result contrasts with the findings of the previously reviewed study conducted by Boxell and others, where the level of affective polarization in the United Kingdom is not very high and the trend is decreasing. The theoretical tool *Action-guiding mental states* of our notion of polarization enables us to explain why this behavior is closely linked to having a certain level of confidence in the core beliefs of the group one identifies with. Giving less amount of money to participants that belong to a political party of the "other side" is conceptually linked to having certain degree of belief in certain beliefs. Crucially, according to our notion, this type of tool is better positioned to measure affective polarization when the level is not yet so high and therefore goes unnoticed for tools such as the feeling thermometer. It is not about what participants say, but also about the trust they exhibit to have in those on opposing parties and in the in-group.

There are at least three recent studies focused on the Spanish case in relation to affective polarization. One of them is the recent study conducted by Miller and

Torcal (Miller & Torcal 2020). In their study, they obtained the data from three different international studies: CSES, the Comparative National Elections Project (CNEP) and the E-DEM project, which include questions about respondents' feelings toward certain political leaders (i.e., a version of the feeling thermometer). So, as in the first study reviewed in this section, the only measure used here is the feeling thermometer, and the affect-based questions are mixed in general surveys with many other different non-affect-based questions.

Two other recent studies focused on Spain are the one conducted by 40dB for EL PAÍS about people's perception of the current situation, and the study conducted by the *Instituto Catalán Internacional para la Paz* (ICIP) about coexistence in Catalonia. These surveys used slightly more indirect questions than those of the feeling thermometer tool to measure affective polarization. In the 40dB survey, for example, some questions read as follows: "When you talk about politics, with whom do you prefer to do so?", "Would you have a drink with a militant of the following political parties?". In the ICIP survey, some questions read as follows: "If the topic of Catalonian independence appears in the following spaces, would you be willing to join the conversation? Conversation with neighbors, At work, Conversation with friends, Conversation with family, Social networks". Arguably, some of these questions are less general than those associated with the feeling thermometer tool and, in that sense, might be a little more useful to measure the participants' level of radicalism when it is still not too high.

However, as in most of the studies reviewed here, the questions aimed at measuring affective polarization are combined with other types of questions. For example, the survey conducted by 40dB also includes direct questions such as "How would you describe the political debate in Spain?", "Who do you think contributes most to the deterioration of the political debate?", "And which political party do you think contributes most to this deterioration?", along with questions about whether respondents believe that the left/right is intolerant, immoral, unpatriotic, undemocratic, or antidemocratic. The ICIP's survey, similarly, also includes direct questions such as "To what extent do you think Catalonian society is polarized according to the following criteria?" (opinions on feminism, immigration, territorial conflict, etc.), "In terms of polarization, how would you rate the degree of polarization in the following areas? Society in general, political parties, the media, myself". Mixing direct questions about the current situation in general, the values of certain parties, etc., with standard questions of affective polarization might have an undesirable effect on the answers given by the respondents, considering the context-sensitivity of the meaning participants express through their responses.

As we have seen in chapter 4, according to several empirical studies (see Napier & Luguri 2016), the use of abstract (vs. concrete) terms is closely related to polarization. But this effect regarding the increase or decrease of polarization crucially depends on other factors. On the one hand, it seems that when conservatives and liberals think in abstract terms, polarization decreases with respect to their biases toward certain social groups (Napier & Luguri 2016: 146-152). On the other hand, it seems that if liberals and conservatives are induced to think in abstract terms and to evaluate some contemporary political issues, polarization increases when partisan identity is salient, while it decreases when a common identity is salient (Napier & Luguri 2016: 153-155). These results are in line with other findings according to which when a common identity is salient, it decreases the degree to which people from one party regard people from other parties under a bad light (Levendusky 2018). However, when a common identity is emphasized, it increases the level of hatred and disapproval toward people forced to leave their home countries (Wojcieszak & Garrett 2018).

Therefore, mixing different types of issues and questions might work as a reminder to participants that we all have some general values in common. As a consequence, it might lower affective polarization on certain questions. And it can have the opposite effect, depending on the issues involved. So, given the highly context-sensitivity of the information expressed through the participants' responses to the tools employed to measure polarization, although it may be more costly and less attractive at first glance, our notion of polarization advices that it is best to avoid mixing these questions in surveys designed to measure polarization.

Finally, we want to close this section by noting that some of the questions included in the surveys aimed at measuring ideological polarization, or to measure what people believe about certain issues, can also measure not what people self-report about their mental states, but what they express, i.e., their commitments especially linked to action, more so when they are compared with participants' responses to other questions. In particular, participants sometimes express their level of attachment to certain core beliefs of a particular identity, i.e., radicalism –what is measured by attitude polarization. For instance, consider the following situation. Between 2007 and 2011, the responses to some opinion's surveys in Spain about the Constitution and the territorial distribution significantly changed. The preference for "A state with a single Central Government, without autonomies", associated with the extreme right of the ideological spectrum, went from being supported by 8.6% of the population (CIS, study 2736, September 2007) to 24.9% (CIS, study 2966, November 2012). As expected, one might conclude from this that there was a shift in people's political beliefs. However, this conclusion is weakened in light of the following observation. A 2016 survey, in addition to the usual questions about support for this or that constitutional reform, reported that almost 50% of respondents claim not to have read any part of the Constitution. 33.3% claimed to have read it partially, and only 15.5% said they had read it in its entirety. Besides, and arguably inconsistently, according to the same study, 77.7% of respondents considered that there is no need to reform the Constitution to give more self-government to Catalonia or to change the territorial model. In this case, the rise of preferences for a state with a single Central Government and rejecting to reform the Constitution to change the territorial model, when most respondents have not read the Constitution, might simply express the rise of radicalism, not extremism. It is not that people claim to believe one thing and actually believe another, but that it is not clear that they believe anything; they are simply expressing, through these responses, the kind of attitudes especially linked to action that are associated with giving great credence to the core beliefs of the group they identify with.

## 8.5. A sketched design of polarization in attitudes

In this section, we briefly offer some recommendations that follow from our notion of polarization, in order to device a design aimed at measuring affective polarization. Recall that our notion of polarization in attitudes recommends that polarization be measured taking into consideration that polarization is highly context-sensitive, and by paying attention to what people express through their evaluative judgments in specific situations, which express the affective attitudes, those attitudes especially linked to action, that are closely connected with having a certain level of radicalism.

In order to comply with the recommendations to measure polarization that follow from the discussion of this dissertation (see section 8.1 and 8.2), many different methodologies can be followed. For instance, the methodology known as 'discussion group' can be employed, which consists of a group of individuals who gather either formally or informally to discuss about a certain topic. The discussions are recorded, transcribed and analyzed qualitatively afterwards. Or some corpus linguistics methodologies can be used to analyze quantitatively and qualitatively the language used by groups of people previously identified as sympathetic to certain ideologies. Or the methodology of vignettes can be employed. In line with this third possibility, we propose as a possible design to describe short narratives and specific situations capturing an active political topic at a given time, and then ask participants to evaluate the narrative or situation. Through their responses, participants will presumably make evaluative judgments that express their level of confidence in an ideological position about the particular topic, and hence their level of radicalism.

In order to develop this sketched design, the first step would be to identify the positions on certain topics that are salient for certain political identities at a particular moment, i.e., the core beliefs of certain political parties and identities. To do so, one option is to track the central political issues in a given society. The surveys of the *Centro de Investigaciones Sociológicas* (CIS) is another source that allows to know which are the hot political topics in Spain, because it incorporates questions periodically repeated about it. This step can be done in a different and automated way, for instance by making use of some digital humanities tools to find out a set of topics associated with a particular political party or ideology. The Latent Dirichlet Analysis (LDA), for example, is one of the available digital tools of Topic Models for text analysis that allows to make a representation of the topics of a corpus. In particular, LDA is a generative probabilistic model that assumes that each topic is a mixture of a set of words, and then is used to determine the topics behind a document, as well as the relevance of each one. This tool can be used on transcripts of parliamentary debates or similar texts, such as the Minutes of the Sessions of the Spanish Parliament,[4] and focus the search on the interventions from a particular political party, or on several political Twitter accounts. After the analysis, we have to select some of the most relevant topics collected, e.g., immigration, monarchy, inclusive language, free speech, false reports, etc.

The second step would be to devise a short narrative counting a story, or to describe a particular situation in which someone says or does something, that represents a particular position on the selected topic. The representation of the position on that topic can be achieved through different and complementary ways. For instance, the narrative or the description of a situation could contain a set of words and expressions mostly associated with the discourse of a particular political party or ideology. In the case of the population forced to leave their home countries, for example, some of the expressions commonly used by right-wing parties are 'illegal', 'invasion', 'danger', 'border assault', etc. So, if we want to elaborate a narrative or situation that captures the core beliefs of a certain right-wing political ideology on this topic, the narrative or description could include these words. But besides, the story of the narrative itself can also represent the position on the topic. For instance, the narrative might be about a public institution allocating funds to help disadvantaged people regardless of their nationality or allocating resources to rescue people who endanger their lives to escape misery in their home countries. Or it could be about a public institution that prioritizes people's nationality when providing certain social aids, etc.

Finally, the last step would be to design certain questions through which participants express their evaluative judgments on the particular narrative or situation. These questions might include terms, expressions and verbs commonly used

[4] This tool has been recently employed by some scholars from the University of Granada as a previous step for a project aimed at tracking the correlation between the presence of situations of crossed disagreement and the rise of polarization. This project has found a very strong correlation between both things.

to evaluate, 'Do you feel rejection / sympathy. . . ?', 'Is it right / wrong. . . ?', 'Would you recommend / warn...?'. 'Would you prefer...?', etc. But the important thing is to make sure that the question indeed demands an evaluation of a specific situation that implies the expression of respondents' perspective on that topic, the expression of affective attitudes, and this will depend on the particular vignette and the particular question in a particular moment. Let us see some raw and possible examples.

**Vignette 1**

This possible vignette is devised to measure the level of confidence in certain beliefs, which can be seen as possible core beliefs of certain right-wing ideologies at a particular time. These core beliefs are that there must be unrestricted freedom of expression, and that the population forced to leave their home countries should be deported to their countries of origin.

> **Vignette 1.1.**: Hueruelia is a small Spanish city. As a citizen of Hueruelia, you can say whatever you want whenever you want without anyone sanctioning you. In addition, illegal people without Spanish nationality are not allowed to enter and are persecuted and deported.

We have introduced the label 'illegal people' in the vignette because, presumably, it will trigger more sympathy or aversion toward these ideas. Then, after reading such a vignette, participants would have to respond some questions that require them to make an evaluative judgment about Hueruelia. For example, one question might be something like this: If a beloved relative had to move to another city to live for a long time, would you recommend (or warn) him/her to move to Hueruelia? Rank your response from 0 (Not at all) to 7 (Absolutely). Other possible question: Would you recommend a friend to change his or her job if he or she had to move to Hueruelia for work? Rank your response from 0 (Not at all) to 7 (Absolutely). Arguably, through their responses, participants would express their affective attitudes, which would give us information about their level of confidence (or rejection) in the beliefs behind the vignette. If someone says that she would strongly recommend to her beloved relative to move to Hueruelia, then she would be expressing the kind of affective attitude closely linked to having a high level of confidence in these beliefs, and then a high level of confidence in the core beliefs of certain right-wing ideologies, even if she does not explicitly self-ascribe those beliefs or does not claim to have warm feelings toward the political ideology that has those beliefs as core ones. In this sense, this vignette would serve to measure polarization in attitudes in an early stage, taking into account the philosophical assumptions of our notion (see section 8.1). Consider the following two alternatives of this vignette.

> **Vignette 1.2.**: Hueruelia is a small Spanish city. Despite the fact that in Hueruelia you can't say whatever you want whenever you want, entry is restricted to illegal people without Spanish nationality, who are persecuted and deported.
>
> **Vignette 1.3.**: Hueruelia is a small Spanish city. Despite the fact that Hueruelia does not restrict the entry of illegal people without Spanish nationality, you can say whatever you want whenever you want.

In these two versions of the vignette, we have changed the relative weight of each topic. In the city described in the vignette 1.2. there is no unrestricted freedom of expression, and in the vignette 1.3. the population forced to leave their home countries are welcome and are not persecuted nor deported to their countries of origin. In both vignettes, both issues are posed as a cost. Then, the responses between both vignettes would give us information to discriminate which of these two beliefs the participants have a higher level of confidence in. If someone would recommend more strongly the city described in vignette 1.2. rather than the city described in 1.3., then she would be expressing greater sympathy for the belief that the population forced to leave their home countries should be deported to their countries of origin than for the other belief. If each of these beliefs were a core belief of two different ideologies, then the participant would be expressing more attachment toward the ideology that has, as a core belief, the idea that the population forced to leave their home countries should be deported to their countries of origin. Recall that participant's responses do not necessarily show what the participant actually believes; their responses might simply express certain affective attitudes closely connected with certain level of credibility in the core beliefs of an ideology, and therefore express certain level of attachment toward certain ideology.

These vignettes are just very general and crude examples that, with further development, could work to measure polarization in attitudes. Ideally, the vignettes should be developed enough so that the questions require as little reflection as possible on what one believes, and then the answers should try to capture the respondents' commitments to the position represented in the vignette. Apart from the development of vignettes with similar characteristics to the ones we have proposed, one of the things that follows from this thesis is that, in order to measure polarization more accurately and at an early stage, further attention should be paid to indirect analyses such as corpus analysis, liar's games, etc.

## 8.6. Conclusion

Let us return to the Harper's letter's case. Now we can say that to the extent that resistance to changes in mentality aimed at redistributing social power –presented under the form of resistance not to such fair changes, but to an alleged coercion and limitation of our freedom of expression– is one of the central workhorses of a political identity, it can be stated that at least some of the signatories of Harper's letter are, or were, polarized in attitudes. They belong to highly privileged nodes of our socio-normative space, where their previous ability to say offensive things without reprisal was not a consequence of free speech, but simply of their unfair privilege. In fact, it is not only that freedom of speech remains intact if socially powerful people are penalized for making publicly offensive statements, but that actual freedom of speech increases, insofar as the offensive actions, verbal and nonverbal, taken by the powerful are themselves a barrier to socially disadvantaged people being able to make use of their freedom of expression. Recently, the University of North Carolina in the United States rescinded an offer to Tenure to Hannah-Jones because of the backlash of conservatives concerned about her involvement in The Times Magazine's 1916 project, which examined the legacy of slavery in America. As Ta-Nehisi Coates recently said on UNC-TV about this case, this is an instance of a general ”cancel policy”. Cancelation as public policy is what many of those who now complains about ”the dictatorship of political correctness” and the ”cancel culture” have been doing since long time ago. Harper's letter seems to contribute to this kind of cancelation as a policy carried out by socially privileged people. Advocating that freedom of expression is being undermined because certain powerful people receive social sanctions after making offensive statements, thereby ignoring all the evidence against this position, is a way of expressing radicalism, i.e., polarization in attitudes. In other words, it is a way of expressing the affective attitudes connected with having a high level of confidence in the core beliefs of certain ideology.

As we have said, polarization in attitudes is the concept of polarization that we have proposed as a result of reassessing the concept of affective polarization. This proposed concept conceives polarization mainly as different versions of radicalism, where the main point is to have a high level of confidence in the core beliefs of one's political group, which leads to be impervious to the reasons and arguments provided by the opposite political side. The tools commonly employed to measure affective polarization are effective in measuring the degree of imperviousness because they involve evaluative language, and through them respondents frequently express their perspective, i.e., their level of attachment to certain topics and political ideologies. We have proposed that, in order to measure polarization in attitudes, the employed tools have to be designed to measure the attitudes that people express, rather than those they self-report. In particular, people's commitments have to be measured in an indirect way, for example by asking them to evaluate a particular situation, i.e., to judge a specific situation that represents a particular position on a certain topic. This concept of polarization satisfies the five requirements we have put forward for a suitable concept of polarization and is free from the objections to the concept of affective polarization.

# Chapter 9

# Summary and Final Notes on Depolarization

In this dissertation, we have proposed a reassessment of the concept of affective polarization. We have argued that the difference between the concepts of ideological polarization and affective polarization does not lie in that the former has to do with political beliefs while the latter deals with people's feelings toward in-group and out-group people, as it is commonly assumed in the literature. Rather, the difference lies in that affective polarization has to do with the *degree of belief*, while ideological polarization has to do with *belief contents*. In particular, affective polarization has to do with the level of confidence in the core beliefs of the political group that people identify with, while ideological polarization is conceived in terms of the distance and other parameters related to the location of certain belief contents in an ideological spectrum. According to this diagnosis, people are affectively polarized when they have a high level of confidence in certain beliefs, leading them to become impervious to others, i.e., to disregard the arguments that come from the other side, regardless of whether their beliefs are located at the center or near of an extreme of the ideological spectrum. The credence in certain beliefs is an affective attitude, an attitude especially linked to certain courses of action. To differentiate our reassessment of the concept of affective polarization from the way it is commonly conceived, we have called our proposal ‘polarization in attitudes’.

Polarization in attitudes is characterized by taking into account the highly context-dependent nature of polarization and by rejecting the first-person authority thesis. The topics that are central to certain political identities are highly sensible to the particularities of different societies, and can also vary from time to time within the same society. For example, one topic might be central to right-wing political identities in Spain but not in the United Kingdom, and might be a core belief of certain identities in 2019 but not in 2021. Moreover, the concept of polarization in attitudes satisfies the five desiderata proposed in this dissertation for a suitable concept of polarization, namely: DANGEROUSNESS, EVIDENCE, RATIONALITY, DISANALOGY and INTERVENTION, only with the help of the philosophical assumptions introduced in chapters 6 and 7. First, the high level of confidence in the core beliefs of a certain political identity, in a particular moment, permits to explain the link between the rise of polarization and the decrease of democratic quality. Second, this high level of confidence in certain beliefs, i.e., radicalism, is quite consistent with the best available evidence on how we polarize. Third, the understanding of polarization behind this concept puts the focus on structural elements of our informational environment rather than on individual factors, avoiding the irrationality story of polarization. Fourth, our approach to mental state attributions and the evaluative use of language enables us to accommodate the disanalogy between self-reporting a mental state and the expression of the mental state in which one actually is, and then to avoid the concerns related to the first-person authority thesis. Finally, our diagnosis enables us to intervene as soon as possible by devising novel ways of measuring polarization to detect it at an early stage.

In addition to taking into consideration the central topics for certain ideological identities, in that particular moment, among which polarization is going to be measured, we have made three recommendations in order to measure polarization. First, people’s mental states must be measured in an indirect way, attending not to what people say about their own states of mind, but to the mental states that they express to be in through what they say and do. In particular, the relevant mental states to measure polarization in attitudes are those especially linked to action, those that give information about what can be expected from people that express those attitudes, and that are closely connected with having certain level of radicalism. Second, our tools to measure polarization must involve evaluative uses of language, because through them we tend to express our practical commitments, i.e., to express our mental states, our world-picture, our attitudes connected with our level of radicalism. Finally, participants' responses must be as specific as possible, i.e., their answers should be judgments of a specific situation, and not claims about general issues. Of course, some of these recommendations serve not only to measure polarization in attitudes, but also ideological polarization. That is, if we want to know the belief content of a group of people rather than their degree of belief, we need to measure the mental states that people express to be in, and not those people self-report.

Our concept of polarization in attitudes rests to a large extent on our discussion of, and more specifically on our approach to, two philosophical assumptions. These assumptions are the first-person authority thesis and the sharp distinction between belief-like and desire-like mental states. We have rejected both. On the one hand, we have argued that having a mental state of the type of beliefs and desires is to have certain conceptual commitments linked to certain course of action and, in that sense, there is not a sharp distinction between belief-like and desire-like mental states in relation with their motivational component. On the other hand, we have argued that in many cases, especially in those that involve complex issues as in the case of political polarization, there is a deep gulf between the mental states in which we say we are in, and those we actually are in, and that we express through the evaluative use of language and certain behavior. In this sense, there is no first-person authority. The dispositional view of mental ascriptions, inspired by some of Wittgenstein's insights, introduced in chapter 6, and the semantic theory introduced in chapter 7, allow us to reassess the concept of affective polarization as we have done in chapter 8.

To close this dissertation, we summarize what we have done in each chapter in section 9.1, discuss a little bit two theoretical tools sometimes used as weapons to foster polarization, that is, the phenomenon of crossed disagreements and the phenomenon of recurring debates (section 9.2), and review certain strategies to depolarize in section 9.3.

## 9.1. Results

In this section, we summarize the conclusions we have achieved in each chapter of this dissertation. In chapter 2, we have introduced a state of the art of the concept of political polarization in terms of political beliefs. To do so, we have payed especial attention to different concepts of polarization that can be distinguished in the literature. In particular, we have differentiated between three categories: forms, types and understandings of polarization. All forms of polarization are conceptually compatible with each other and with all types of polarization. All types of polarization can be conceived in terms of all understandings of polarization, except the type of polarization "ideological polarization" that is essentially characterized by the understanding of polarization that conceives it in terms of belief

In chapter 3, we have introduced the concept of affective polarization, commonly conceived as one not having to do with beliefs, but with people's feelings. The aim of this chapter was twofold. First, we discussed some limitations of the concepts of ideological and affective polarization, as they are commonly understood in the literature. Second, we made explicit two challengeable philosophical assumptions behind both concepts of polarization: the first-person authority thesis and the sharp distinction between belief-like and desire-like mental states regarding their link to action. Taken together, these limitations placed both concepts in a bad position. After that, we introduced a group of conditions, five desiderata, that we think a concept of polarization must meet in order to be an operational one. Finally, we suggested that both concepts of polarization, as they are commonly understood, cannot meet such desiderata, and briefly outlined how our concept of affective polarization reassessed, i.e., polarization in attitudes, can meet those desiderata.

In chapter 4, we have examined the best available evidence concerning how we get polarized. The main goal of this chapter was to review the body of evidence that a suitable concept of polarization, an operational one, must be compatible with. In particular, we have payed especial attention to analyze whether this body of evidence is more compatible with one or another understanding of polarization of those introduced in chapter 2. We have shown that radicalism, the understanding of polarization that conceives it in terms of degree of belief, is better positioned than the understanding of polarization in terms of belief content to accommodate the evidence. In this sense, we have suggested that the suitable notion of polarization should be conceived in terms of radicalism rather than in terms of belief content. Besides, we have made an argument regarding the rational story of the rise of polarization: taken together, all the evidence concerning how we get polarized is compatible with the idea that polarization is the result of a rational process.

In chapter 5, we have explored whether descriptivist views on mental ascriptions can provide the adequate philosophical framework for a suitable concept of polarization. We have argued that most of descriptivist positions cannot satisfy the desideratum DISANALOGY, and have held that those descriptivist positions that don't endorse first-person authority and therefore seem to be able to satisfy DISANALOGY, cannot accommodate certain relevant evidence, more specifically our intuitions as competent speakers triggered by different cases of belief self-ascriptions where, sometimes, the speaker does exhibit authority. Besides, these positions cannot explain those cases where, through a mental self-ascription, the speaker does not describe a particular state of affairs that could be or not the case but also expresses something else: she expresses certain information of what can be expected from her, she expresses some of her attitudes beyond the belief self-ascribed. Thus, these positions seemed to be bad positioned to meet EVIDENCE.

In chapter 6, we have provided a pragmatist and nondescriptivist approach to mental ascriptions, based on some of Wittgenstein's insights, which is compatible with the desiderata proposed for a suitable concept of polarization. In particular, one key idea of this position is that there is possibility of error regarding mental ascriptions. This approach opened two possibility of errors. First, someone might self-ascribe a belief or another mental state and not being in that mental state. Second, someone might self-ascribe a belief or another mental state and, through it, express some different attitudes beyond that particular belief. More specifically, one might express certain attitudes closely linked to action, the kind of information conceptually connected with what can be expected from her. The view offered in this chapter can explain why these cases are possible: the meaning expressed through our claims is highly context-dependent and is underpinned by the norms governing our social practices. This view enables us to satisfy the desiderata DISANALOGY and EVIDENCE, and puts us on the right path to meet the other ones.

In chapter 7, we have complemented the view introduced in chapter 6 but from a different and important angle. In particular, we have argued that, through the evaluative use of language, we usually express our commitments, our attitudes especially linked to certain courses of action, and not just those commitments or mental states that we self-report to have. We started by introducing, from an intuitive perspective, the distinction between the descriptive and the evaluative. Then, we introduced expressivism, a semantic theory that accommodates this difference particularly well. Crucially, we have argued that not every sort of expressivism can do the work here: only those expressivisms compatible with the view introduced in chapter 6, that is, those expressivisms that don't entail an internalist approach to certain mental states and assume the possibility of error in ascribing mental states, can explain why through the evaluative use of language one can express her attitudes, her level of radicalism.

In chapter 8, we completed the argument of this dissertation: we offered our notion of polarization in attitudes as a result of a reassessment of the notion of affective polarization, and explained how it meets the proposed desiderata for a suitable concept of polarization. Our notion crucially depends on the philosophical tools introduced in chapters 6 and 7. This notion allows us to explain why some recent studies aimed at measuring polarization might not be measuring what they try to measure, and enables us to offer specific recommendations to measure and capture certain processes of polarization that, otherwise, pass unnoticed. Besides, we have discussed the difference between our notion of polarization and some of the attitudes usually discussed, from the field of epistemology, in relation with the rise of polarization.

## 9.2. Weaponized phenomena: Crossed disagreements and recurrent debates

As introduced in section 4.5, crossed disagreements are situations where two parts disagree on a certain topic and both display signs of conceiving the disagreement in significantly different terms, for instance one part conceives it as a factual discussion and the other as a normative one. Situations of this kind, especially when take place in public contexts, are potentially pernicious to democracy because they increase the size and density of the pool of arguments each side is exposed to, and as a consequence the audience ends up reinforcing their initial positions, becoming more polarized than at the beginning. Crossed disagreements are potentially dangerous when systematically appear in public debates with certain objectives, in particular to advance certain political agendas. But it is important to note that they are not necessarily pernicious: a crossed disagreement might be used as a strategy of resistance, for instance by moving a discussion which is factual to a normative domain and then being able to discuss something that had been initially assumed and that was a pernicious assumption.

One way this phenomenon is sometimes used to advance certain political agendas in public debates is through a subtle process. It can be argued that certain public figures with a privileged background, such as Donald Trump and Boris Johnson, cultivate a public persona, of someone who is reckless and irresponsible, for whom the conditions of being member of a disenfranchised identity group apply, at least for the eyes of their supporters: they systematically receive less credibility than other public figures because they belong to the group of people who say what they think and are persecuted by "political correctness". This constructed image of someone unreliable, perceived as an injustice by their supporters, is systematically used to take advantage of public debates to generate situations of crossed disagreements, and then obtain political benefits from it (see Almagro et al. forthcoming).

Another phenomenon that sometimes is used, consciously or unconsciously, as a weapon to increase the level of polarization, and that is to some extent similar to crossed disagreements, can be called *recurrent debates.* This phenomenon consists of bringing back debates on the nature or pertinence of well-established democratic values or newly recognized rights in a recurrent manner with the aim of casting doubts about the facts supporting them and, in the end, eroding their status as full-blown rights or values (Almagro & Heras-Escribano ms). This phenomenon, then, can be seen as the strategy of taking descriptive statements supporting certain rights or values and making them look like evaluative statements in an unjustified and covert way, simply by putting them up for debate in a *recurrent manner*. This way, it will look like a particular right, value or principle is never fully recognized or granted in its entirety if a certain amount of the population is constantly casting doubts about it. This is used as a strategy for polarizing given the repetitive nature of the phenomenon, that contributes to increase the pool of arguments that people are exposed to, and not taking seriously different societal groups to which certain rights have been recognized, which undermines their status as full-blown citizens. Thus, this erodes the genuine deliberative nature of democracy. It also happens that certain people even try to camouflage, consciously or unconsciously, this undermining strategy under the right to freedom of expression, and when they are singled out for undermining the rights of certain groups they complain that freedom of expression has been restricted, as was the case with Harper's letter.

Both mechanisms can be seen as weapons to foster polarization in attitudes to the extent that both contribute to rise the level of confidence in the core beliefs of their respective groups.

## 9.3. Depolarization

Finally, we want to close this work by saying something very briefly about the issue of depolarization. Admittedly, even though depolarization is an important topic related to the one of political polarization, it is not the subject of this work, and for that reason the comments we are going to make about it are very limited.

David Adler has recently found that, contrary to what is commonly assumed in the literature about political polarization, pernicious attitudes toward democracy are more strongly held not by those who self-identify with a political identity close to the political extremes, but by those who self-identify as centrists (Adler 2018). This finding coheres with our diagnosis in this dissertation: the type of polarization that endangers democracy does not have necessarily to do with having certain belief contents or positions located near of the extremes of an ideological spectrum, but with the level of confidence in the core beliefs of the political identity that people identify with, no matter where these beliefs are located in the political spectrum. Perceiving oneself as centrist, or even apolitical, is also an ideological and political identity, and thus one can be also polarized even if one self-identify as apolitical or centrist. So, depolarization should be aimed at decreasing some people's level of confidence in certain beliefs.

What are the strategies available to carry out this undertaking? What should these intervention strategies look like? As we have seen in chapter 4, getting exposed to the other side's arguments can polarize (see also Mutz 2006 for a review). Communication across lines of difference can pose potential dangers, in particular it can increase our level of confidence in our prior beliefs. Besides, it has been argued that both priming partisan ambivalence and promoting apolitical mechanisms for maintaining a good self-imagen among polarized people fail in depolarizing (Levendusky 2018). In other words, asking polarized people, with the aim to depolarize, to reflect, for instance, on what they dislike about their own political group and like about the opponent political group, and to reflect on their nonpolitical virtues, does not really decrease their level of polarization. That is, trying to bring positions closer in this way does not seem to work.

Moreover, trying to show people on the other side that they are wrong on factual issues is neither very effective. In fact, this usually has the opposite effect: it makes the other person become more polarized, or at least leaves her political preferences unaltered. That is, at best, fact-checking only allows the other side to become more accurate in their factual argumentation, remaining their initial political preferences for the most part intact (Porter et al. 2019). Then, the successful strategies designed to depolarize have to take all this into consideration.

As we have argued along this dissertation, the level of credibility in the core beliefs of the political group that we identify with is tied to certain practices, that is, it is closely linked to the things we do and the things that can be reasonably expected from us. There is a normative link between our level of confidence in the core beliefs of our group, the things we believe, say and do. So, the intervention strategy should be directed at modifying our practices, i.e., it must be structural. As Appiah points out, "And when it comes to change, what moves people is often not an argument from a principle, not a long discussion about values, but just a gradually acquired new way of seeing things" (Appiah 2007: 152). Besides, the required interventions have to be designed to cover a wide variety of social scenarios. On the one hand, the interventions aimed at depolarizing can be directed at changing the architecture of our digital environment, in particular the dynamics that foster polarization. On the other hand, the intervention should also address phenomena such as crossed disagreements and recurring debates, as well as the use of certain language in public domains (see section 9.2). And, of course, these strategies can be complemented with other more individual interventions, as, for instance, the policy related to our revision responsibility (see section 8.3.2). In what follows, we will briefly review some proposed strategies of intervention to depolarize.

Regarding the individual interventions, the findings of a recent study conducted by Abeywickrama and Laham show that when people is asked to advocate for their own opinions, those who experience low confidence during their attempt to argue in favor of their position are more likely to depolarize (Abeywickrama & Laham 2020). It seems that when people realize that they have not so strong reasons as they thought to support their position, they adopt a more receptive attitude. This finding is in line with the phenomenon of illusion of explanatory depth, introduced in section 3.4.1: when we discover that we know less about something than we thought, we tend to decrease our level of confidence on it. So, these findings suggest that rather than trying to convince the other part that your position is the correct one and they are wrong, a more effective strategy in order to decrease of the level of confidence of someone might be to try to force them to advocate for their position. But, again, this strategy will be more or less successful depending on the context.

In relation to our current online system, which seems to be mainly designed to capture our attention rather than to promote deliberation and autonomous choices, and thus contributes to the rise of polarization, some authors (Lorenz-Spreen et al. 2020) have recently proposed to use the behavioral sciences. In particular, certain technological cues to indicate the epistemic quality of online contents, the factors underlying algorithmic decisions, and the degree of consensus in online debates, and harness these cues to design two types of behavioral interventions –nudging (see Sunstein 2008) and boosting (see Kozyreva et al. 2020)– to redesign online environments for informed and autonomous choice, and therefore depolarize. Nudging and boosting are two strategies for intervention that have been proven to be effective in different domains (Arno & Thomas 2020; Kurvers et al. 2016; Lusardi & Mitchell 2014). The nudging strategy proposed consists in altering the online environment so as to draw users' attention to these cues, and the boosting strategy consists in teaching people to search information attending to the relevant cues. But which are these cues?

These authors distinguish between endogenous and exogenous cues. Endogenous cues refer to the content itself, like the plot or the actors and their relations (Lorenz-Spreen et al. 2020: 2). Modern search engines use natural language-processing tools that analyze content, and can accomplish this objective. However, it is not yet sufficiently sophisticated and presents considerable problems in order to being able to indicate the epistemic quality of a content (Lorenz-Spreen et al. 2020: 2). Exogenous cues, on the other hand, refer to the context of the information, and an example would be the Google's PageRank algorithm. Authors focus on exogenous cues and in how they can be harnessed to facilitate intervention. As an example of endogenous clues that might highlight the epistemic quality of individual articles serve the following: a newspaper article's sources and citations, reference to established concepts and empirical evidence, and objectivity of the language.

Thus, an example of nudging intervention would be to change the choice architecture online by adding certain information, collected by the previous clues, such as highlighting when the content comes from anonymous sources, contextualizing the number of likes and shares by expressing them against the absolute frequency of total readers, showing the average reading time, etc. On the other hand, an example of boosting intervention would be to increase the possibility of customizing users' news feed, in which each item is transparently accompanied by the relevant information about its epistemic quality, or to foster the competence for distinguish between high and low quality sources with fast-and-frugal decision trees, where the user is guided to scrutinize relevant cues to select epistemically good information (Lorenz-Spreen et al. 2020: 5).

In relation to phenomena fostering polarization that take place in public contexts, such as crossed disagreements, a possible intervention would be, for example, to implement policies of intervention that do not only take into consideration the available time to each part, but also monitoring the speeches that, consciously or unconsciously, are aimed at generating a crossed disagreement. In that sense, the person in charge of moderating a debate could not only stop an intervention when time is running out, but also make explicit those movements that are taking place in a more subtle way with the aim to reframe the debate.

Regarding the phenomenon of recurrent debates, where factual statements that, in virtue of their being factual deserve the status of being assumptions, is presented, in a recurrent and covert manner, as a non-factual claim, undermining thus its deserved character of assumption, a possible intervention would be to avoid entering into the alleged debate. That is, as a move aimed to cancel the invitation to question the factual nature of a claim, a policy would be to avoid consider that claim as one for which there is an ongoing debate. In other words, instead of try to argue why the other part is wrong, a better option would be to avoid entering the debate, because otherwise we run the risk of generating a crossed disagreement, or giving the impression that the issue is indeed an open question.

To the extent that these phenomena are related to the rise of polarization, to intervene in them can count as a way to depolarize, because we reduce the contexts where, as a consequence of discussing and judging together, we get polarized. But a lot of work is still needed: we have to devise new intervention strategies to improve our current divided condition.

# Conclusiones

En esta sección, resumimos las conclusiones a las que hemos llegado en cada capítulo de esta tesis doctoral. En el capítulo 2, hemos introducido el estado de la cuestión del concepto de polarización política en términos de creencias políticas. Para ello, hemos prestado especial atención a los diferentes conceptos de polarización que se pueden distinguir en la literatura. En particular, hemos diferenciado tres categorías: formas, tipos y concepciones de la polarización. Todas las formas de polarización son conceptualmente compatibles entre sí y con los tipos de polarización. Todos los tipos de polarización, por su parte, pueden concebirse bajo cualquiera de las dos concepciones de polarización que hemos distinguido, excepto el tipo de polarización "polarización ideológica", que se caracteriza esencialmente por la concepción de la polarización que la concibe en términos de contenido de creencias y, por definición, no puede concebirse por tanto en términos de radicalismo.

En el capítulo 3, hemos introducido el concepto de polarización afectiva, comúnmente concebido como el tipo de polarización que tiene que ver no con las creencias, sino con los sentimientos de una población. El objetivo de este capítulo fue doble. En primer lugar, discutimos algunas limitaciones de los conceptos de polarización ideológica y afectiva, tal y como se entienden habitualmente en la literatura. En segundo lugar, hemos hecho explícitos dos asunciones filosóficas cuestionables de ambos conceptos de polarización: la tesis de la autoridad de la primera persona y la distinción radical con respecto a su vínculo con la acción entre los estados mentales similares a las creencias y aquellos similares a los deseos. En conjunto, estas limitaciones dejan a ambos conceptos en una mala posición. A continuación, introdujimos un grupo de condiciones, cinco desiderata, que pensamos que debe cumplir un concepto de polarización para ser operativo. Por último, sugerimos que ambos conceptos de polarización, tal y como se entienden comúnmente en la literatura, no pueden cumplir esos desiderata, y esbozamos brevemente cómo nuestro concepto de polarización afectiva reevaluada, es decir, la polarización en actitudes, puede cumplir esos desiderata.

En el capítulo 4, hemos examinado la mejor evidencia de la que disponemos sobre cómo nos polarizamos. El objetivo principal de este capítulo fue revisar el conjunto de evidencia con la que debe ser compatible un concepto adecuado y operativo de polarización. En particular, hemos prestado especial atención a analizar si este conjunto de evidencia es más compatible con una u otra concepción de polarización de las dos que introdujimos en el capítulo 2. Hemos mostrado que el radicalismo, la concepción de la polarización que la entiende en términos de grado de creencia, está mejor posicionada que la concepción de la polarización que la entiende en términos de contenido de creencia para acomodar la evidencia. Hemos sugerido que la noción adecuada de polarización debería concebirse en términos de radicalismo y no en términos de contenido de creencia. Además, hemos presentado un argumento a favor de la explicación racional del aumento de la polarización: en conjunto, toda la evidencia relativa a cómo nos polarizamos es compatible con la idea de que la polarización es el resultado de un proceso racional.

En el capítulo 5, hemos explorado si las aproximaciones descriptivistas a las adscripciones mentales pueden proporcionar el marco filosófico adecuado para un concepto apropiado de polarización. Hemos argumentado que la mayoría de las posiciones descriptivistas no pueden satisfacer el desiderátum DISANALOGÍA, y hemos sostenido que aquellas posiciones descriptivistas que no se comprometen con la autoridad de la primera persona y que, por tanto, parecen poder satisfacer la DISANALOGÍA, no pueden dar cabida sin embargo a cierta evidencia relevante, más concretamente no pueden dar cuenta de nuestras intuiciones como hablantes competentes ej diferentes casos de auto atribución de creencia en los que, en ocasiones, el hablante sí exhibe autoridad. Además, estas posiciones no pueden explicar aquellos casos en los que, a través de una auto atribución mental, el hablante no describe un estado de cosas particular que podría ser o no el caso, sino que (también) expresa algo más: expresa cierta información de lo que se puede esperar de ella, expresa algunas de sus actitudes más allá de la creencia auto atribuida. Por lo tanto, estas posiciones parecen estar mal posicionadas para cumplir con el desiderátum EVIDENCIA.

En el capítulo 6, hemos proporcionado un enfoque pragmatista y no descriptivista de las atribuciones mentales, basado en algunas de las ideas de Wittgenstein, que es compatible con los desiderata propuestos para un concepto adecuado de polarización. En particular, una idea clave de esta posición es que existe la posibilidad de error respecto a las atribuciones mentales. Este enfoque abre dos posibilidades de error. En primer lugar, alguien podría auto atribuirse una creencia u otro estado mental y no estar en ese estado mental. En segundo lugar, alguien podría auto atribuirse una creencia u otro estado mental y, a través de tal auto atribución, expresar algunas actitudes diferentes a la creencia auto atribuida particular. Más concretamente, uno podría expresar ciertas actitudes especialmente vinculadas a la acción, el tipo de información conceptualmente relacionada con lo que se puede esperar de quien hace la auto atribución. La aproximación que se ofrece en este capítulo puede explicar por qué estos casos son posibles: el significado expresado a través de nuestras afirmaciones depende en gran medida del contexto y se determina en virtud de las normas que rigen nuestras prácticas sociales. Esta aproximación nos permite satisfacer los desiderata DISANALOGÍA y EVIDENCIA, y nos pone en el camino correcto para satisfacer los demás.

En el capítulo 7, hemos complementado la aproximación introducida en el capítulo 6, pero desde un ángulo diferente e importante para los propósitos de esta tesis. En particular, hemos argumentado que, a través del uso evaluativo del lenguaje, solemos expresar nuestros compromisos, nuestras actitudes especialmente vinculadas con ciertos cursos de acción, y no solo aquellos compromisos o estados mentales que nos auto atribuimos. Comenzamos introduciendo, desde un punto de vista intuitivo, la distinción entre lo descriptivo y lo evaluativo. A continuación, hemos introducido el expresivismo, una teoría semántica que sirve especialmente bien para dar cuenta de esta diferencia. De manera crucial, hemos argumentado que no todo tipo de expresivismo puede hacer el trabajo aquí: solo aquellos expresivismos compatibles con la aproximación introducida en el capítulo 6, es decir, aquellos expresivismos que no implican una posición internalista de ciertos estados mentales y que por tanto asumen la posibilidad de error en la atribución de estados mentales, pueden explicar por qué a través del uso evaluativo del lenguaje uno puede expresar sus actitudes, su nivel de radicalismo.

En el capítulo 8, hemos completado el argumento de esta tesis doctoral: hemos ofrecido nuestra noción de polarización en actitudes como resultado de una reevaluación de la noción de polarización afectiva, y hemos explicado cómo esta noción cumple los desiderata propuestos para un concepto adecuado de polarización. Nuestra noción depende crucialmente de las herramientas filosóficas introducidas en los capítulos 6 y 7. Esta noción nos permite explicar por qué algunos estudios recientes destinados a medir la polarización podrían no estar midiendo lo que intentan medir, y nos permite ofrecer recomendaciones específicas para medir y capturar ciertos procesos de polarización que, de otro modo, pasan desapercibidos. Además, hemos discutido la diferencia entre nuestra noción de polarización y algunas de las actitudes habitualmente discutidas, desde la epistemología, en relación con el aumento de la polarización.

# Bibliography

(2017).

Abeywickrama, R. S. & Laham, S. M. (2020). Meta-cognition predicts attitude depolarization and intentions to engage with the opposition following pro-attitudinal advocacy. *Social Psychology*, 51(6), 408–421.

Abramowitz, A. (2010). *The disappearing center: Engaged citizens, polarization, and American democracy*. Yale University Press.

Abramowitz, A. I. (2006). Comment on disconnected: The political class versus the people. *Red and blue nation*, (pp. 72–84).

Abramowitz, A. I. (2007). Constraint, ideology, and polarization in the american electorate: Evidence from the 2006 cooperative congressional election study. In *Annual Meeting of the American Political Science Association, Chicago, IL* (pp. 634–652).

Abramowitz, A. I. & Saunders, K. L. (2008). Is polarization a myth? *The Journal of Politics*, 70(2), 542–555.

Abramowitz, A. I. & Stone, W. J. (2006). The bush effect: Polarization, turnout, and activism in the 2004 presidential election. *Presidential Studies Quarterly*, 36(2), 141–154.

Abramowitz, A. I. & Webster, S. W. (2017). Taking it to a new level: Negative partisanship, voter anger and the 2016 presidential election. In *State of the Parties Conference* Akron, Ohio: University of Akron.

Acero, J. J. & Villanueva, N. (2012). Wittgenstein y la intencionalidad de lo mental. *Análisis Filosófico*, XXXII(2), 117–154.

Adler, D. R. (2018). The centrist paradox: Political correlates of the democratic disconnect.

Aikin, S. F. & Talisse, R. B. (2020). *Political argument in a polarized age: Reason and democratic life*. Polity Press.

Albarracín, D., Johnson, B. T., & Zanna, M. P. (2005). *The handbook of attitudes*. Lawrence Erlbaum Associates Publishers.

Allen, T. (2020). *Engaging Others*. PhD thesis, University of Connecticut, Connecticut.

Almagro, M., Hannikainen, I. R., & Villanueva, N. (Forthcoming). Whose words hurts? contextual determinants of offensive speech. *Personality and Social Psychology Bulletin*.

Almagro, M. & Heras-Escribano, M. (ms). Recurring debates.

Almagro, M. & Moreno, A. (Forthcoming). Affective polarization and testimonial and discursive injustice. In D. Bordonaba, V. Fernández, & J. R. Torices (Eds.), *The Political Turn in Analytic Philosophy. Reflections on Social Injustice and Oppression*. De Gruyter.

Almagro, M., Navarro-Laespada, L., & Pinedo, M. (2022). Is testimonial injustice epistemic? let me count the ways. *Hypatia. A Journal of Feminist Philosophy*.

Almagro, M., Osorio, J., & Villanueva, N. (2021). Weaponized testimonial injustice. *Las Torres de Lucca*.

Almagro, M. & Villanueva, N. (2021). Polarización y tecnologías de la información: radicales vs. extremistas. *Dilemata. International Journal of Applied Philosophy*, 34(13), 51–69.

Almond, G. A. & Verba, S. (1963). *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton Legacy Library.

Appiah, K. A. (2007). *Cosmopolitanism: Ethics in a World of Strangers*. Norton and Company.

Applebaum, A. (2020). *Twilight of Democracy The Seductive Lure of Authoritarianism*. Doubleday.

APSA (1950). Part i. the need for greater party responsibility. *The American Political Science Review*, 44(3), 15–36.

Arno, A. & Thomas, S. (2020). The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systematic review and meta-analysis. *BMC Public Health*, 16.

Arvan, M. (2019). The dark side of morality: Group polarization and moral epistemology. In *The Philosophical Forum*, volume 50 (pp. 87–115).: Wiley Online Library.

Ayala, S. (2016). Speech affordances: A structural take on how much we can do with our words. *European Journal of Philosophy*, 24(4), 879–891.

Ayala, S. (2018). A structural explanation of injustice in conversation: It's about norms. *Pacific Philosophical Quarterly*.

Ayer, A. J. (1936/2001). *Language, Truth and Logic*. London: Penguin.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 115 (pp.3̃7).

Baker, L. (2007). Social externalism and first-person authority. *Erkenntnis*, 67, 287–300.

Bar-On, D. (2004). *Speaking my mind: Expression and Self-Knowledge*. Oxford University Press.

Bar-On, D. (2015). Sociality, expression, and this thing called language. *Inquiry: An Interdisciplinary Journal of Philosophy*, 59(1), 56–79.

Bar-On, D. (2019). Neo-expressivism: (self-)knowledge, meaning and truth. In M. J. Frápolli (Ed.), *Expressivisms, Knowledge and Truth* (pp. 11–34). Royal Institute of Philosophy.

Bar-On, D. & Chrisman, M. (2009). Ethical neo-expressivism. *Oxford Studies in Metaethics*, 4, 132–165.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, (pp. 1–12).

Baron, R. S., Hoppe, S. I., Feng, C., Brunsman, B., Linneweh, B., & Rogers, D. (1996). Social corroboration and opinion extremity. *Journal of Experimental Social Psychology*, 32, 537–560.

Barroso, P. M. (2015). *Grammar, Expressiveness, and Inter-subjective Meanings: Wittgenstein's Philosophy of Psychology.* Cambridge Scholars Publishing.

Barz, W. (2018). Is there anything to the authority thesis? *Journal of Philosophica Research*, 43, 125–143.

Battaly, H. (2021). Closed-mindedness and arrogance. In A. Tanesini & M. P. Lynch (Eds.), *Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives* (pp. 53–70). Routledge.

Battich, L., Fairhurst, M., & Deroy, O. (2020). Coordinating attention requires coordinated senses. *Psychonomic Bulletin and Review*, 27, 1126–1138.

Bayne, T. & Hattiangadi, A. (2013). Belief and its bedfellows. In N. Nottelmann (Ed.), *New Essays on Belief* (pp. 124–144). Palgrave Macmillan.

Bayrakly, E. & Hafez, F. (2020). *European Islamophobia Report 2019.* Istambul.

Bilgrami, A. (2006). *Self-Knowledge and Resentment.* Harvard University Press.

Billig, M. & Tajfel, H. (1973). Social categorization and similarity in intergroup behaviour. *European Journal of Social Psychology*, 3(1), 27–52.

Bishop, B. (2008). *The big sort: Why the clustering of like-minded America is tearing us apart.* Houghton Mifflin Harcourt.

Blankenhorn, D. (2015). Why polarization matters. *The American Interest.*

Boghossian, P. A. (1989). Content and self-knowledge. *Philosophical Topics*, 17(1), 5–26.

Bordonaba, D. (2017). *Higher-Order Operators and Taste Predicates: An Expressivist Proposal.* PhD thesis, University of Granada, Granada.

Bordonaba, D. (2020). Los peligros de las cámaras de eco. *Endoxa*, 45, 249–260.

Bordonaba, D., Fernández, V., & Torices, J. R. (2022). *The Political Turn in Analytic Philosophy.* De Gruyter.

Bordonaba, D. & Villanueva, N. (2018). Affective polarization as impervious reasoning. In *Philosophical Perspectives. The 13th conference of the Italian Society for Analytic Philosophy*: Italian Society for Analytic Philosophy.

Bordonaba, D. & Villanueva, N. (Forthcoming). Retractación y contextualismo: Nuevas condiciones de adecuación. In D. Pérez (Ed.), *Contextualismo Semántico.* Prensas de la Universidad de Zaragoza.

Borgoni, C. (2015a). Dissonance and doxastic resistance. *Erkenntnis*, 80(5), 957–974.

Borgoni, C. (2015b). Dissonance and moorean propositions. *Dialectica*, 69(1), 107–127.

Borgoni, C. (2016). Dissonance and irrationality: A criticism of the in-between account of dissonance cases. *Pacific Philosophical Quarterly*, 97(1), 107–127.

Borgoni, C. (2018a). Authority and attribution: the case of epistemic injustice in self-knowledge. *Philosophia.*

Borgoni, C. (2018b). Unendorsed beliefs. *Dialectica*, 72(1), 49–68.

Borgoni, C. (Forthcoming). First-person authority and the social aspects of self-knowledge. In J. Lackey & A. McGlynn (Eds.), *Oxford Handbook of Social Epistemology*. Oxford University Press.

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2020). *Cross-country trends in affective polarization*. Technical report, National Bureau of Economic Research.

Bramson, A., Grim, P., Singer, D. J., Berger, W. J., Sack, G., Fisher, S., Flocken, C., & Holman, B. (2017). Understanding polarization: meanings, measures, and model evaluation. *Philosophy of science*, 84(1), 115–159.

Brentano, F. (1874). *Psychology from an empirical standpoint*. Routledge.

Breton, A. & Dalmazzone, S. (2002). Information control, loss of autonomy, and the emergence of political extremism. *Political extremism and rationality*, (pp. 44–66).

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin*, 17(5), 475–482.

Brewer, M. D. (2005). The rise of partisanship and the expansion of partisan conflict within the american electorate. *Political Research Quarterly*, 58(2), 219–229.

Broncano-Berrocal, F. & Carter, J. A. (2021). *The Philosophy of Group Polarization: Epistemology, Metaphysics, Psychology*. Routledge.

Brown, R. (1985). *Social Psychology: The Second Edition*. New York: The Free Press.

Bullock, J. G., Gerber, A. S., Hill, S. J., & Hubert, G. A. (2015). Partisan bias in factual beliefs about politics. *Quarterly Journal of Political Science*, 10, 519–578.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, 4(1), 73–121.

Burnstein, E. & Vinokur, A. (1977). Persuasive argumentation and social comparison as determinants of attitude polarization. *Jouranl of Experimental Social Psychology*, 13, 315–332.

Byrne, A. (2018). *Transparency and Self-Knowledge.* Oxford University Press.

Carlin, R. E. & Love, G. J. (2018). Political competition, partisanship and interpersonal trust in electoral democracies. *British Journal of Political Science*, 48(1), 115–139.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), 284–299.

Carmona, C. & Villanueva, N. (ms). Situated judgments as a new model for intercultural communication.

Carothers, T. & O'Donohue, A. (2019). *Democracies divided: The global challenge of political polarization.* Brookings Institution Press.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge.* Oxford University Press.

Cassam, Q. (2019). *Vices of the Mind: From the Intellectual to the Political.* Oxford: Oxford University Press.

Cassam, Q. (2021). The polarisation toolkit. In A. Tanesini & M. P. Lynch (Eds.), *Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives* (pp. 212–228). Routledge.

Cepollaro, B., Soria, A., & Stojanovic, I. (2021). The semantics and pragmatics of value judgments. In P. Stalmaszczyk (Ed.), *The Cambridge Handbook of Philosophy of Language* chapter 24. Cambridge University Press.

Cerezo, M. (1998). *Lenguaje y lógica en el Tractatus.* Universidad de Navarra.

Cerezo, M. (2003). Isomorfismo y proyección en el tractatus. In J. J. Acero (Ed.), *Viejos y nuevos pensamientos.* Comares.

Charlow, N. (2014). The problem with the Frege-Geach problem. *Philosophical Studies*, 167(3), 635–665.

Chen, H., Reardon, R., Rea, C., & Moore, D. J. (1992). Forewarning of content and involvement: Consequences for persuasion and resistance to persuasion. *Journal of Experimental Social Psychology*, 28, 523–541.

Child, W. (2017). The inner and the outer. In H.-J. Glock & J. Hyman (Eds.), *A Companion to Wittgenstein* (pp. 465–477). Blackwell.

Chrisman, M. (2007). From epistemic contextualism to epistemic expressivism. *Philosophical Studies*, 135(2), 225–254.

Clutton, P. (2018). A new defence of doxasticism about delusions: The cognitive phenomenological defence. *Mind and Language*, (pp. 1–20).

Coliva, A. (2016). *The varieties of self-knowledge*. Palgrave Macmillan.

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. E. Apter (Ed.), *Ideology and Discontent*. New York: The Free Press of Glencoe.

Crary, A. & Read, R. (2000). *The New Wittgenstein*. New York: Routledge.

Dain, E. (2019). Wittgenstein on belief in other minds.

Darwall, S. L. (2006). *The Second-Person Standpoint*. Harvard University Press.

Davidson, D. (1984). First person authority. *Dialectica*, 38(2-3), 101–111.

Davidson, D. (1987). Knowing ones own mind. *Proceedings and Addresses of the American Philosophical Association*, 60(3), 441–458.

Davies, A. (2020). Identity display: another motive for metalinguistic disagreement. *Inquiry: An Interdisciplinary Journal of Philosophy*, (pp. 1–23).

De Bruin, L. & Strijbos, D. (2020). Does confabulation pose a threat to first-person authority? mindshaping, self-regulation and the importance of self-know-how. *Topoi: An International Review of Philosophy*, 39, 151–161.

De Cruz, H. & De Smedt, J. (2013). The value of epistemic disagreement in scientific practice. the case of homo floresiensis. *Studies in History and Philosophy of Science Part A*, 44(2), 169–177.

De Mesel, B. (2019). Are moral judgments semantically uniform? a wittgensteinian approach to the cognitivism - non-cognitivism debate. In B. De Mesel & O. Kuusela (Eds.), *Ethics in the Wake of Wittgenstein* (pp. 126–148). New York: Routledge.

Deroy, O. (2019). The danger of bursting bubbles.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have american's social attitudes become more polarized? *American journal of Sociology*, 102(3), 690–755.

Dorst, K. (2020). Reasonably polarized: Why politics is more rational than you think.

Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82(1), 62–68.

Downey, D. J. & Huffman, M. L. (2001). Attitudinal polarization and trimodal distributions: measurement problems and theoretical implications. *Social science quarterly*, 82(3), 494–505.

Downs, A. (1957). An economic theory of political action in a democracy. *Journal of Political Economy*, 65(2), 135–150.

Druckman, J. N. & Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1), 114–122.

Drury, M. O. (1973). *The Danger of Words*. Routledge.

Eagly, A. H. & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College.

Egan, A. (2010). Disputing about taste. In T. Warfield & R. Feldman (Eds.), *Disagreement* (pp. 247–286). Oxford University Press.

Eidenmuller, M. E. (2008). Transcription of obama's speech 2004 democratic national convention keynote address.

El Pueblo, P. (2020). Hay una gran mafia intentando lucrarse con denuncias falsas contra maestros.

Ellis, C. & Stimson, J. A. (2012). *Ideology in America.* New York: Cambridge University Press.

Evans, G. (1982). *The Varieties of Reference.* Oxford University Press.

Falvey, K. (2000). The basis of first-person authority. *Philosophical Topics*, 28(2), 69–99.

Fantl, J. (2018). *The Limitations of the Open Mind.* Oxford: Oxford University Press.

Fernández, V. & Heras-Escribano, M. (2020). Social cognition: a normative approach. *Acta Analytica*, 35(1), 75–100.

Fernbach, P., Rogers, T., Cragi, F., & Sloman, S. (2013). Political extremism is supported by an illusion of understanding. *Association for Psychological Science*, 24(6), 939–946.

Ferrer, J. (2014). El papel de la segunda persona en la constitución del autoconocimiento. *Diamon*, 62, 71–86.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7(2), 117–140.

Field, H. (2009). Epistemology without metaphysics. *Philosophical Studies*, 143(2), 249–290.

Field, H. (2018). Epistemology from an evaluativist perspective. *Philosophers' Imprint*, 18(2).

Finkelstein, D. (2003). *Expression and the Inner.* Harvard University Press.

Fiorina, M. P. (2017). *Unstable Majorities: Polarization, Party Sorting, and Political Stalemate.* Hoover Press.

Fiorina, M. P., Abrams, S. A., & Pope, J. C. (2008). Polarization in the american public: Misconceptions and misreadings. *The Journal of Politics*, 70(2), 556–560.

Fiorina, M. P. & Abrams, S. J. (2008). Political polarization in the american public. *Annual Review of Political Science*, 11, 563–588.

Fiorina, M. P. & Levendusky, M. S. (2006). Disconnected: The political class versus the people. *Red and blue nation*, 1, 49–71.

Fisher, M., Goddu, M. K., & Keil, F. (2015). Searching for explanations: How the internet inflates estimates of internal knowledge. *Journal of Experimental Psychology*, 144(3), 674–687.

Fodor, J. (1979). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences*, 3(1), 63–73.

Fogelin, R. (1985). The logic of deep disagreements. *Informal Logic*, 7(1), 3–11.

Fomina, J. (2019). Of "patriots" and citizens: Asymmetric populist polarization in poland. In T. Carothers & A. O'Donohue (Eds.), *Democracies Divided: The Global Challenge of Political Polarization* (pp. 126–150). Brookings Institution Press.

Forero-Mora, J. A. & Frápolli, M. J. (2021). Show me. tractarian non-representationalism. *Teorema*, XL(2), 1–19.

Forro, J. (1987). *Psychosemantics*. Cambridge University Press.

Frápolli, M. J. (2019). Introduction: Expressivisms, knowledge and truth. In M. J. Frápolli (Ed.), *Expressivisms, Knowledge and Truth* (pp. 1–9). Cambridge: Cambridge University Press.

Frápolli, M. J. (2019). The pragmatic gettier: Brandom on knowledge and belief. *Disputatio*, 8(9), 563–591.

Frápolli, M. J. (2019). Propositions first: Biting Geach's bullet. In M. J. Frápolli (Ed.), *Expressivisms, Knowledge and Truth* (pp. 87–110). Cambridge: Cambridge University Press.

Frápolli, M. J. & Romero, E. (2003). *Meaning, Basic Self-Knowledge, and Mind*. CSLI Publications.

Frápolli, M. J. & Villanueva, N. (2012). Minimal expressivism. *Dialectica*, 66(4), 471–487.

Frápolli, M. J. & Villanueva, N. (2015). Expressivism, relativism, and the analytic equivalence test. *Frontiers in Psychology*, 6.

Frápolli, M. J. & Villanueva, N. (2016). Pragmatism. propositional priority and the organic model of propositional individuation. *Disputatio*, 46, 203–217.

Frápolli, M. J. & Villanueva, N. (2018). Minimal expressivism and the meaning of practical rationality. In *Rationality and Decision Making* (pp. 1–22). Brill Rodopi.

Fricker, M. (2007). *Epistemic Injustice. Power and the Ethics of Knowing*. Oxford University Press.

Geach, P. T. (1960). Ascriptivism. *The Philosophical Review*, 69(2), 221–225.

Gendler, T. S. (2008). Alief in action (and reaction). *Mind and Language*, 23(5), 552–585.

Gentzkow, M. (2016). Polarization in 2016.

Gibbard, A. (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge (Mass.): Harvard University Press.

Gibbard, A. (2003). *Thinking How to Live*. Cambridge (Mass.): Harvard University Press.

Gibbard, A. (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

Gidron, N., Adams, J., & Horne, W. (2020). *American Affective Polarization in Comparative Perspective*. Cambridge University Press.

Gilovich, T. (1991). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. New York: The Free Press.

Gomila, A. (2002). La perspectiva de la segunda persona de la atribución mental. *Azafea*, 4, 123–138.

González, J. J. & Bouza, F. (2009). *Las razones del voto en la España democrática 1977-2008.* Catarata.

Gottfried, J. & Grieco, E. (2018). Younger americans are better than older americans at telling factual news statements from opi- nions. Pewresearch.

Groenendijk, J. & Stokhof, M. (1991). Dynamic Predicate Logic. *Linguistics and Philosophy*, 14(1), 39–100.

Grossmann, M. & Hopkings, D. A. (2016). *Asymmetric politics: Ideological Republicans and group interest Democrats.* Oxford University Press.

Hacker, J. & Pierson, P. (2005). *Off-Center: The Republican Revolution and the Erosion of American Democracy.* New Haven: Yale University Press.

Hacker, P. M. S. (2005). Of knowledge and knowing that someone is in pain. In A. Pichler & S. Saatela (Eds.), *Wittgenstein: The Philosopher and His Works.* The Wittgenstein Archives at the University of Bergen.

Hallson, B. G. & Kappel, K. (2020). Disagreement and the division of epistemic labor. *Synthese*, 197, 2823–2847.

Hallsson, B. G. (2019). The epistemic significance of political disagreement. *Philosophical Studies*, 176(8), 2187–2202.

Hare, C., Armstrong, D. A., Bakker, R., Carroll, R., & Poole, K. (2015). Using bayesian aldrich-mckelvey scaling to study citizens' ideological preferences and perceptions. *American Journal of Political Science*, 59(3), 759–774.

Hare, C., McCarty, N., Poole, K., & Rosenthal, H. (2012). Polarization is real and asymmetric. Project Voteview Blog.

Hare, R. (1952). *The Language of Morals.* Oxford University Press.

Harman, G. (1973). *Thougth.* Princeton University Press.

Harteveld, E. & Wagner, M. (2020). Affective polarization across parties: why do people dislike some parties more than others?

Haslanger, S. (2015). Distinguished lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy*, 45(1), 1–15.

Hastorf, A. H. & Cantril, H. (1954). They saw a game: A case study. *The Journal of Abnormal and Social Psychology*, 49(1), 129–134.

Heltzel, G. & Laurin, K. (2020). Polarization in america: two possible futures. *Current Opinion in Behavioral Sciences*, 34, 179–184.

Heras-Escribano, M. & Pinedo, M. (2016). Are affordances normative? *Phenomenology and the Cognitive Sciences*, 15(4), 565–589.

Hetherington, M. J. (2001). Resurgent mass partisanship: The role of elite polarization. *American Political Science Review*, (pp. 619–631).

Hetherington, M. J. (2009). Putting polarization in perspective. *British Journal of Political Science*, (pp. 413–448).

Hetherington, M. J., Long, M. T., & Rudolph, T. J. (2016). Revisiting the myth: New evidence of a polarized electorate. *Public Opinion Quarterly*, 80, 321–350.

Hetherington, M. J. & Rudolph, T. J. (2015). *Why Washington won't work: Polarization, political trust, and the governing crisis*, volume 104. University of Chicago Press.

Hetherington, M. J. & Weiler, J. D. (2009). *Authoritarianism and polarization in American politics*. Cambridge University Press.

Houston, D. A. & Fazio, R. H. (1989). Based processing as a function of attitude accessibility: Making objective judgments subjectively. *Social Cognition*, 7(1), 51–66.

Huddy, L., Mason, L., & Aarøe, L. (2015). Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review,*, 109(1), 1–17.

Hume, D. (1739/2007). *A Treatise of Human Nature*. Clarendon Press.

Hurlburt, R. T. & Schwitzgebel, E. (2007). *Life and mind: Philosophical issues in biology and psychology. Describing inner experience? Proponent meets skeptic.* MIT Press.

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Jouurnal of Personality and Social Psychology*, 50(6), 1141–1151.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22, 129–146.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideologya social identity perspective on polarization. *Public Opinion Quarterly*, 76(3), 405–431.

Iyengar, S. & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, 59(3), 690–707.

Jamieson, K. H. & Cappella, J. N. (2010). *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment.* Oxford University Press.

Johnston, C. D., Newman, B. J., & Velez, Y. (2015). Ethnic change, personality, and polarization over immigration in the american public. *Public Opinion Quarterly*, 79(3), 662–686.

Kahan, D. M. (2017). The expressive rationality of inaccurate perceptions. *Behavioral and Brain Sciences.*

Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, 14(2), 147–174.

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *The Journal of Philosophy*, 105(10), 611–633.

Khoo, J. (2017). Code words in political discourse. *Philosophical Topics*, 45(2), 33–64.

Kidd, I. J. (2021). Martial metaphors and argumentative virtues and vices. In A. Tanesini & M. P. Lynch (Eds.), *Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives* (pp. 25–38). Routledge.

Klein, E. (2017). Obamaism sought strength in unity. trumpism finds power through division. Vox.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes*, 56, 28–55.

Kölbel, M. (2004). Faultless disagreement. *Proceedings of the Aristotelian Society*, 104(1), 53–73.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Association for Psychological Science*, 21(3), 103–156.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language.* Cambridge (Mass.): Harvard University Press.

Kukla, R. (2014). Performative force, convention, and discursive injustice. *Hypatia. A Journal of Feminist Philosophy*, 29(2), 440–457.

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology*, 53(4), 636–647.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498.

Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 113 (pp. 8777–8782).

Lasersohn, P. (2005). Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy*, 28(6), 643–686.

Lauka, A., McCoy, J., & Firat, R. B. (2018). Mass partisan polarization: Measuring a relational concept. *American Behavioral Scientist*, 62(1), 107–126.

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, 80, 392–410.

Levendusky, M. (2009). *The partisan sort: How liberals became Democrats and conservatives became Republicans.* University of Chicago Press.

Levendusky, M. (2013). Why do partisan media polarize viewers? *American Journal of Political Science*, 50(3), 611–623.

Levendusky, M. S. (2018). When efforts to depolarize the electorate fail. *Public Opinion Quarterly*, (pp. 1–10).

Levitsky, S. & Ziblatt, D. (2018). *How democracies die.* Broadway Books.

Levy, N. (2019). Due deference to denialism: Explaining ordinary people's rejection of established scientific findings. *Synthese*, 196(1), 313–327.

Levy, N. (2020). Arrogance and servility online: Humility is not the solution. In M. Alfano, M. P. Lynch, & A. Tanesini (Eds.), *The Routledge Handbook on the Philosophy of Humility* (pp. 472–483). Routledge.

Lewis, D. (1979). Attitudes *de dicto* and *de se*. *The Philosophical Review*, 88(4), 513–543.

Lewis, D. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74(4), 549–567.

Liberman, A. & Chaiken, S. (1992). Defensive processing of personality relevant health messages. *Personality and Social Psychology Bulletin*, 18, 669–679.

Loewer, B. & Rey, G. (1991). *Meaning in Mind: Fodor and His Critics.* Oxford University Press.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109.

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., & Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behavior.*

Lupu, N. (2015). Party polarization and mass partisanship: A comparative perspective. *Political Behavior*, 37(2), 331–356.

Lusardi, A. & Mitchell, O. S. (2014). The economic importance of financial literacy: theory and evidence. *Journal of Economic Literature*, 52(1), 5–44.

Lynch, M. P. (2010). Epistemic circularity and epistemic incommensurability. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social Epistemology.* Oxford: Oxford University Press.

Lynch, M. P. (2016). *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data.* Liveright Publishing.

Lynch, M. P. (2019). *Know-it-all society: Truth and arrogance in political culture.* Liveright Publishing.

Lynch, M. P. (2021). Polarisation and the problem of spreading arrogance. In A. Tanesini & M. P. Lynch (Eds.), *Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives* (pp. 141–157). Routledge.

Macarthur, D. (2010). Wittgenstein and expressivism. In D. Whiting (Ed.), *The Later Wittgenstein on Language* (pp. 81–95). Palgrave.

Macdonald, C. (1995). Externalism and first-person authority. *Synthese*, 194, 99–122.

MacFarlane, J. (2014). *Assessment Sensitivity: Relative Truth and Its Applications.* Oxford: Oxford University Press.

Mackie, D. M. (1986). Social identification effects in group polarization. *Journal of Personality and Social Psychology,*, 50(4), 720–728.

Mann, T. & Ornstein, N. (2012). *It's Even Worse than It Looks: How the American Constitutional System Collided with the New Politics of Extremism.* New York: Basic Books.

Manne, K. (2020). *Entitled. How Male Privilege Hurts Women.* Crown.

Maravall, J. A. (1981). Los apoyos partidistas en españa: Polarización, fragmentación y estabilidad. *Revista de Estudios Políticos*, 23, 9–32.

Marcus, E. (2019). Reconciling practical knowledge with self-deception. *Mind*, 128(512), 1205–1225.

Marques, T. (2018). Retractions. *Synthese*, 195, 3335–3359.

Marques, T. & García-Carpintero, M. (2014). Disagreement about taste: commonality presuppositions and coordination. *Australasian Journal of Philosophy*, 92(4), 701–723.

Martiarena, A. (2021). Ayuso: "cuando te llaman fascista es que estás en el lado bueno".

Mason, L. (2013). The rise of uncivil agreement: Issue versus behavioral polarization in the american electorate. *American Behavioral Scientist*, 57(1), 140–159.

Mason, L. (2015). "i disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, 59(1), 128–145.

Mason, L. (2018). *Uncivil agreement: How politics became our identity.* University of Chicago Press.

Mastro, D. (2015). Why the media's role in issues of race and ethnicity should be in the spotlight. *Journal of Social Issues*, 71(1), 1–16.

Matthews, R. J. (2007). *The measure of mind: Propositional attitudes and their attribution.* Oxford University Press.

McCarty, N. (2019). *Polarization: What everyone needs to know®*. Oxford University Press.

McConnell, C. R., Brue, S. L., & Flynn, S. M. (2018). *Economics: principles, problems, and policies, 21th.* McGraw-Hill.

McCoy, J. & Somer, M. (2019). Toward a theory of pernicious polarization and how it harms democracies: Comparative evidence and possible remedies. *The Annals of the American Academy of Political and Social Science*, 681(1), 234–271.

McDowell, J. (2008). Response to dreyfus. *Inquiry*, 50(4), 366–370.

McNally, L. & Stojanovic, I. (2017). Aesthetic adjectives. In J. O. Young (Ed.), *Semantics of Aesthetic Judgements* (pp. 17–37). Oxford University Press.

Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and the social imagination.* Oxford University Press.

Mele, A. R. (2001). *Self-Deception Unmasked.* Princeton University Press.

Mendelberg, T. (2001). *The Race Card: Campain Strategy, Implicit Messages, and The Norm of Equality.* Princeton University Press.

Mercier, H. & Sperber, D. (2017). *The Enigma of Reason.* Harvard University Press.

Miller, A. G., McHoskey, J. W., Bane, C. M., & Dowd, T. G. (1993). The attitude polarization phenomenon: Role of response measure, attitude extremity, and behavioral consequences of reported attitude change. *Journal of Personality and Social Psychology*, 64(4), 561–574.

Miller, L. & Torcal, M. (2020). Veinticinco años de polarización afectiva en españa.

Moore, G. E. (1903/1993). *Principia Ethica.* Cambridge: Cambridge University Press.

Moscovici, S. & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of personality and social psychology*, 12(2), 125.

Mounk, Y. (2018). *The people vs. democracy: Why our freedom is in danger and how to save it.* Harvard University Press.

Mouw, T. & Sobel, M. E. (2001). Culture wars and opinion polarization: the case of abortion. *American journal of Sociology*, 106(4), 913–943.

Munro, G. D. & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, 23(6), 636–653.

Mutz, D. C. (2006). *Hearing the Other Side: Deliberative versus Participatory Democracy.* New York: Cambridge University Press.

Myers, D. G. (1975). Discussion-induced attitude polarization. *Human Relations*, 28(8), 699–714.

Myers, D. G. & Bishop, G. D. (1970). Discussion effects on racial attitudes. *Science*, 169(3947), 778–779.

Myers, D. G. & Lamm, H. (1976). The group polarization phenomenon. *Psychological bulletin*, 83(4), 602.

Napier, J. L. & Luguri, J. B. (2016). From silos to synergies: The effects of construal level on political polarization. In P. Valdesolo & J. Graham (Eds.), *Social Psychology of Political Polarization.* Routledge.

Navajas, J., Heduan, F. A., Garrido, J. M., González, P. A., Garbulsky, G., Ariely, D., & Sigman, M. (2019). Reaching consensus in polarized moral debates. *Current Biology*, 29, 4124–4129.

Navarro-Laespada, L. (Forthcoming). Where are ethical properties? predication, location and category mistake. In N. Saras & M. Bella (Eds.), *Women in Pragmatism: Past, Present, and Future*, Women in the History of Philosophy and Sciences. Springer.

Navarro-Laespada, L. & Frápolli, M. J. (2018). Inferentialism, representationalism and moral responsibility. In *Proceedings of the IX Conference of the Spanish Society of Logic, Methodology and Philosophy of Science* (pp. 151–153).

Nguyen, T. (2020). Echo chambers and epistemic bubles. *Episteme*, 17(2), 141–161.

Nickerson, R. S. (1999). How we know -and sometimes misjudge- what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, 125(6), 737–759.

Nisbett, R. & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.

Nuñez de Prado-Gordillo, M. (2022). *Non-descriptivism in mental health and the functional assessment-based approach to delusions.* PhD thesis, Universidad Autónoma de Madrid.

Nyhan, B. & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32, 303–330.

Ogden, C. K. & Richards, I. A. (1923). *The Meaning of Meaning.* Harvest Book.

O'Leary-Hawthorue, J. & Price, H. (1996). How to stand up for non-cognitivists. *Australasian Journal of Philosophy*, 74(2), 275–292.

Olsson, E. J. (2013). A bayesian simulation model of group deliberation and polarization. In *Bayesian argumentation* (pp. 113–133). Springer.

Ortoleva, P. & Snowberg, E. (2015). Overconfidence in political behavior. *American Economic Review*, 105(2), 504–535.

Osorio, J. & Villanueva, N. (2019). Expressivism and crossed disagreements. In M. J. Frápolli (Ed.), *Expressivisms, Knowledge and Truth.* Royal Institute of Philosophy.

Palfrey, T. & Poole, K. (1987). The relationship between information, ideology, and voting behavior. *American Journal of Political Science*, (pp. 511–530).

Pariser, E. (2011). *The Filter Bubble: What the Internet is Hiding from You.* New York: The Penguin Press.

Pérez-Navarro, E. (2019). *Ways of Living*. PhD thesis, University of Granada, Granada.

Pérez-Navarro, E. (2021). The way things go: Moral relativism and suspension of judgment. *Philosophical Studies*.

Pew Research, C. (2019). Partisan antipathy: More intense, more personal.

Pinedo, M. (2004). De la interpretación radical a la fusión de horizontes: La perspectiva de la segunda persona. In J. J. Acero (Ed.), *El Legado de Gadamer*. Universidad de Granada.

Pinedo, M. (2018). From volleying to distributed embodied rationality. In M. Hetmański (Ed.), *Rationality and Decision Making*, volume 111 (pp. 119–138). Poznań Studies in the Philosophy of the Sciences and the Humanities.

Pinedo, M. (2020). Ecological psychology and enactivism: A normative way out from ontological dilemmas. *Frontiers in Psychology*, 11, 1–10.

Pinedo, M. & Villanueva, N. (Forthcoming). Epistemic de-platforming. In D. Bordonaba, V. Fernández, & J. R. Torices (Eds.), *The Political Turn in Analytic Philosophy. Reflections on Social Injustice and Oppression*. De Gruyter.

Plunkett, D. (2015). Which concepts should we use?: Metalinguistic negotiations and the methodology of philosophy. *Inquiry*, 58(7-8), 828–874.

Porter, E., Wood, T. J., & Bahador, B. (2019). Can presidential misinformation on climate change be corrected? evidence from internet and phone experiments. *Research and Politics*, (pp. 1–10).

Price, H. (2011). Expressivism for two voices. In J. Knowles & H. Rydenfelt (Eds.), *Pragmatism, Science and Naturalism*. Peter Lang.

Pylyshyn, Z. (1984). *Computation and Cognition*. Cambridge University Press Cambridge.

Ramos, J. (2001). Confusiones gramaticales acerca de lo mental. In J. J. Botero (Ed.), *El Pensamiento de L. Wittgenstein* (pp. 201–220). Editorial Aula.

Razinsky, H. (2017). *Ambivalence. A Philosophical Exploration.* Rowman and Littlefield.

Reiljan, A. (2020). Fear and loathing across party lines (also) in europe: Affective polarisation in european party systems. *European journal of political research*, 59(2), 376–396.

Richard, M. (1981). Temporalism and eternalism. *Philosophical Studies*, 39(1), 1–13.

Robson, J. (2014). A social epistemology of aesthetics: belief polarization, echo chambers and aesthetic judgement. *Synthese*, 191(11), 2513–2528.

Rogowski, J. C. & Sutherland, J. L. (2016). How ideology fuels affective polarization. *Political Behavior*, 38(2), 485–508.

Rorty, R. (1980). *Philosophy and the Mirror of Nature.* Princeton University Press.

Rozenblit, L. & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, 26, 521–562.

Russell, B. (1986). Theory of knowledge. In *The Collected Papers of Bertrand Russell*, volume Volumen 8: The Philosophy of Logical Atomism and Other Essays 1914-1919. George Allen and Unwin.

Ryle, G. (1949/2009). *The Concept of Mind.* London: Routledge.

Sambrotta, M. & García-Jorge, P. A. (2018). Expressivism without mentalism in meta-ontology. *International Journal of Philosophical Studies*, 26(5), 781–800.

Sánchez-Curry, D. (2018). *How Beliefs Are Like Colors.* PhD thesis, Faculties of the University of Pennsylvania.

Sartori, G. (1976). *Parties and Party Systems: A Framework for Analysis.* New York: Cambridge University Press.

Saul, J. (2018). Dogwhistles, political manipulation, and philosophy of language. In D. Fogal, D. W. Harris, & M. Moss (Eds.), *New Works on Speech Acts* (pp. 360–383). Oxford University Press.

Schachter, S. & Wheeler, L. (1962). Epinephrine, chlorpromazine, and amusement. *The Journal of Abnormal and Social Psychology*, 65(2), 121–128.

Schroeder, M. (2008). *Being For: Evaluating the Semantic Program of Expressivism.* Clarendon Press.

Schwitzgebel, E. (2001). In-between believing. *The Philosophical Quarterly*, 51(202), 76–82.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Nous*, 36(2), 249–275.

Schwitzgebel, E. (2008). The unreliability of naïve introspection. *Philosophical Review*, 117, 245–273.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between ocurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91, 531–553.

Schwitzgebel, E. (2011a). Knowing your own beliefs. *Canadian Journal of Philosophy*, 35, 41–62.

Schwitzgebel, E. (2011b). *Perplexities of Consciousness.* MIT Press.

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New Essays on Belief* (pp. 75–99). London: Palgrave Macmillan.

Shea, N. (2018). *Representation in Cognitive Science.* Oxford: Oxford University Press.

Sides, J., Tesler, M., & Vavreck, L. (2018). *Identity crisis.* Princeton University Press.

Sieber, J. & Ziegler, R. (2019). Group polarization revisited: A processing effort account. *Personality and Social Psychology Bulletin*, 45(10), 1482–1498.

Simpson, M. (2020). What is global expressivism? *The Philosophical Quarterly*, 70(278), 140–161.

Smith, M. (1987). The humean theory of motivation. *Mind*, 96(381), 36–61.

Smith, P. S. & Lynch, M. P. (2020). Varieties of deep epistemic disagreement. *Topoi: An International Review of Philosophy*.

Soria, A. (2019). *The Place of Value in Natural Language*. PhD thesis, Institut Jean Nicod, Paris.

Soria, A. & Stojanovic, I. (2019). On linguistic evidence for expressivism. In M. J. Frápolli (Ed.), *Expressivisms, Knowledge and Truth*, volume 86 (pp. 155–180). Cambridge University Press.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, 25(4), 359–393.

Srinivasan, A. (2015). Normativity without cartesian privilege. *Philosophical Issues*, 25, 273–299.

Srinivasan, A. (2016). Philosophy and ideology. *Theoria*, 31(3), 371–380.

Srinivasan, A. (2020). Radical externalism. *Philosophical Review*, 129(3), 395–431.

Stalnaker, R. (1978). Assertion. In P. Cole (Ed.), *Pragmatics* (pp. 315–332). New York: New York Academy Press.

Stanley, J. (2015). *How Propaganda Works*. Princeton: Princeton University Press.

Stavrakakis, Y. (2018). Paradoxes of polarization: Democracy's inherent division and the (anti-) populist challenge. *American Behavioral Scientist*, 62(1), 43–58.

Sterelny, K. (1990). *The Representational Theory of Mind*. Blackwell.

Stern, D. G. (1995). *Wittgenstein on Mind and Language*. Oxford University Press.

Stevenson, C. (1937). The emotive meaning of ethical terms. *Mind*, 46(181), 14–31.

Sundell, T. (2016). The tasty, the bold, and the beautiful. *Inquiry*, 59(6), 793–818.

Sunstein, C. R. (2002). The law of group polarization. *The Journal of Political Philosophy*, 10(2), 175–195.

Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness.* Yale University Press.

Sunstein, C. R. (2009). *Going to extremes: How like minds unite and divide.* Oxford University Press.

Sunstein, C. R. (2017). *Republic: Divided Democracy in the Age of Social Media.* Princeton: Princeton University Press.

Sunstein, C. R., Schkade, D., Ellman, L. M., & Sawicki, A. (2006). *Are Judges Political? An Empirical Analysis of the Federal Judiciary.* Brookings Institution Press.

Taber, C. S. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755–769.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, 223(5), 96–103.

Tajfel, H., Billig, M., Bundy, R. P., & Flamant, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1(2), 149–178.

Tajfel, H. & Turner, J. C. (1979). An integrative theory of inter-group conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of inter-group relations* (pp. 33–47). Brooks/Cole.

Talisse, R. B. (2019). *Overdoing democracy: Why we must put politics in its place.* New York: Oxford University Press.

Tanesini, A. (2016). I—'calm down, dear': Intellectual arrogance, silencing and ignorance. *Aristotelian Society Supplementary Volume*, 90(1), 71–92.

Tanesini, A. (2021). Arrogance, polarisation and arguing to win. In A. Tanesini & M. P. Lynch (Eds.), *Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives* (pp. 158–174). Routledge.

Tanesini, A. & Lynch, M. P. (2021). *Polarisation, Arrogance, and Dogmatism: Philosophical Perspectives.* Routledge.

Theriault, S. (2013). *The Gingrich Senators.* Oxford: Oxford University Press.

Torices, J. R. (2019). *Ranking The World Through Words: Disagreement, Dogwhistles, and Expressivism.* PhD thesis, University of Granada, Granada.

Torices, J. R. (Forthcoming). Understanding dogwhistle politics. *Theoria.*

Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest.* Yale University Press.

Turner, J. C. (1981). Towards a cognitive redefinition of the social group. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, 1(2), 93–118.

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Association for Psychological Science*, (pp. 1–7).

Valentino, N. A., Hutchings, V. L., Banks, A. J., & Davis, A. K. (2008). Is a worried citizen a good citizen? emotions, political information seeking, and learning via the internet. *Political Psychology*, 29(2), 247–273.

Vekony, R., Mele, A., & Rose, D. (2020). Intentional action without knowledge. *Synthese.*

Velleman, D. (2009). *How We Get Along.* Cambridge University Press.

Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Scientific Reports*, (pp. 1–12).

Viciana, H., Hannikainen, I. R., & Gaitan Torres, A. (2019). The dual nature of partisan prejudice: Morality and identity in a multiparty system. *PloS one*, 14(7).

Villanueva, N. (2014). Know thyself: A tale of two theses and two theories. *Teorema*, 33(3), 49–65.

Villanueva, N. (2019). Wittgenstein: descripciones y estados mentales. In J. J. Acero (Ed.), *Guía Comares de Wittgenstein* (pp. 145–170). Comares.

Vinokur, A. & Burnstein, E. (1978). Depolarization of attitudes in groups. *Journal of Personality and Social Psychology*, 36(8), 872–885.

Vitriol, J. & Marsh, J. (2018). The illusion of explanatory depth and endorsement of conspiracy beliefs. *European Journal of Social Psychology*.

Webster, S. W. & Abramowitz, A. I. (2017). The ideological foundations of affective polarization in the us electorate. *American Politics Research*, 45(4), 621–647.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford University Press.

Westfall, J., Van Boven, L., Chambers, J. R., & Judd, C. M. (2015). Perceiving political polarization in the united states: Party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspectives on Psychological Science*, 10(2), 145–158.

Westwood, S. J., Iyengar, S., Walgrave, S., Leonisio, R., Miller, L., & Strijbis, O. (2018). The tie that divides: Cross-national evidence of the primacy of partyism. *European Journal of Political Research*, 57(2), 333–354.

Wike, R. & Simmons, K. (2015). Global support for principle of free expression, but opposition to some forms of speech. *Pew Research Center, Global Attitudes and Trends*.

Williamson, T. (2002). *Knowledge and its Limits*. Oxford University Press on Demand.

Wilson, T. (2002). *Strangers to ourselves: Discovering the adaptive unconscious.* Harvard University Press.

Wittgenstein, L. (1922/1975). *Tractatus Logico-Philosophicus.* London: Routledge.

Wittgenstein, L. (1953/2009). *Philosophical Investigations.* Oxford: Wiley-Blackwell.

Wittgenstein, L. (1958). *The Blue and Brown Books.* Blackwell.

Wittgenstein, L. (1965). A lecture on ethics. *The Philosophical Review*, 74, 3–12.

Wittgenstein, L. (1967). *Zettel.* Blackwell.

Wittgenstein, L. (1969). *On Certainty.* Blackwell.

Wittgenstein, L. (1974). *Philosophical Grammar.* Blackwell.

Wittgenstein, L. (1980a). *Culture and Value.* Blackwell.

Wittgenstein, L. (1980b). *Remarks on the Philosophy of Psychology, Volume I.* Blackwell.

Wittgenstein, L. (1980c). *Remarks on the Philosophy of Psychology, Volume II.* Blackwell.

Wittgenstein, L. (1992). *Last Writings on the Philosophy of Psychology, vol. II, The Inner and the Outer, 1949-51.* Blackwell.

Wojcieszak, M. & Garrett, R. K. (2018). Social identity, selective exposure, and affective polarization: How priming national identity shapes attitudes toward immigrants via news selection. *Human communication research*, 44(3), 247–273.

Wolfe, A. (1998). *One nation, after all: How middle-class Americans really think about: God, country, family, racism, welfare, immigration, homosexuality, work, the right, the left, and each other.* Viking.

Wood, T. J. & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, 41, 135–169.

Wright, C. (1998). Self-knowledge: The wittgensteinian legacy. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing our own minds* (pp. 13–45). Clarendon Press.

Yalcin, S. (2018). Expressivism by force. In D. Fogal, D. W. Harris, & M. Moss (Eds.), *New Work on Speech Acts* (pp. 400–428). Oxford: Oxford University Press.

Yamagishi, T. (2001). Trust as a form of social intelligence. *American Psychological Association.*

Yamagishi, T., Kikuchi, M., & Kosugi, M. (2002). Trust, gullibility, and social intelligence. *Asian Journal of Social Psychology*, 2(1), 145–161.

Yavorsky, J. E., Kamp, C. H., & Schoppe-Sullivan, S. J. (2015). The production of inequality: The gender division of labor across the transition to parenthood. *Journal of marriage and the family.*

Zajonc, R. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, 10(6), 224–228.