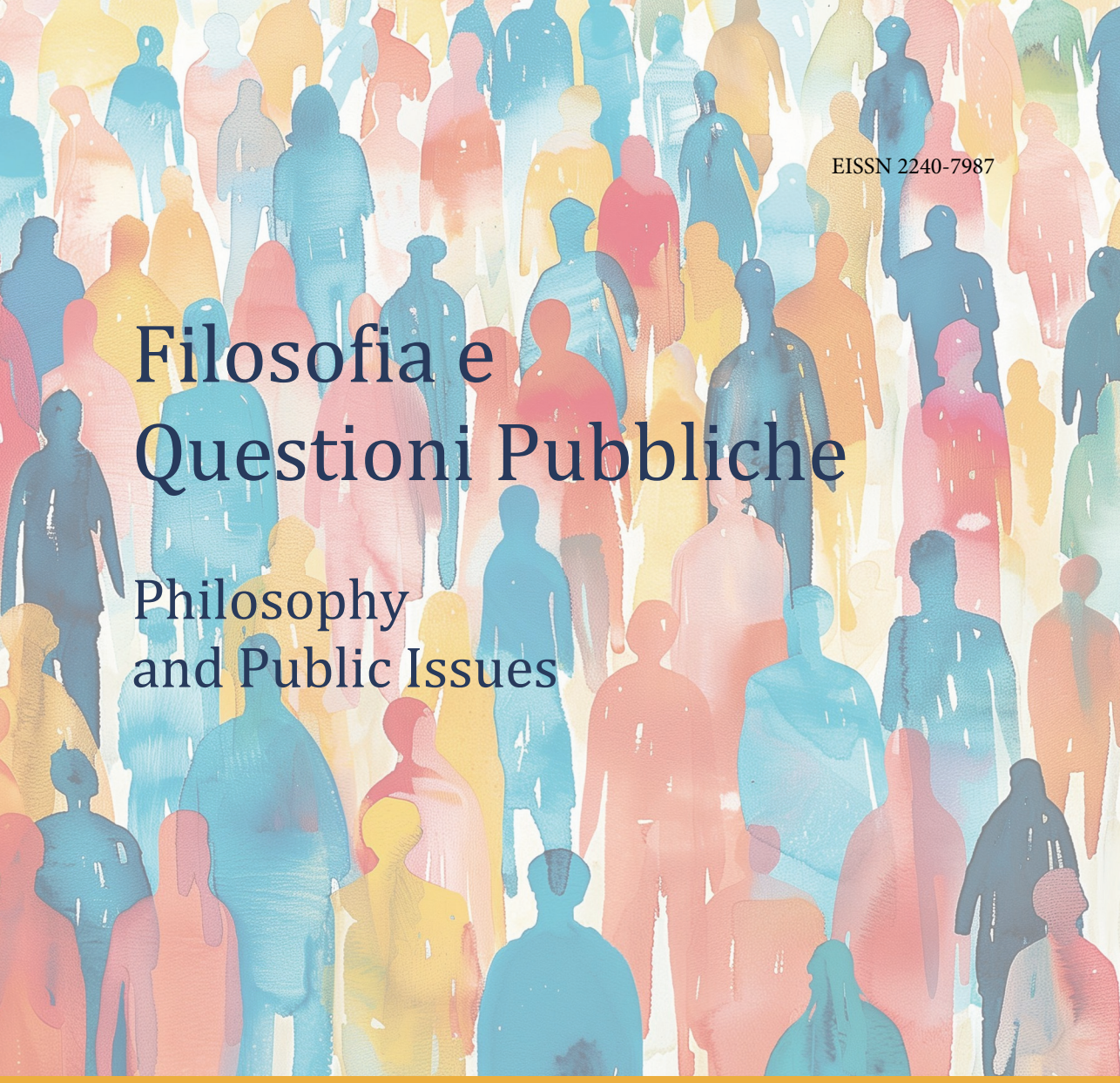
EISSN 2240-7987



Filosofia e Questioni Pubbliche

Philosophy and Public Issues

2/2024

FQP PPI

LUISS

ETHOS
Osservatorio di Etica Pubblica
Luiss Business School

Giappichelli

Filosofia e Questioni Pubbliche

Philosophy and Public Issues

2/2024

FQP PPI

LUISS

ETHOS Osservatorio di Etica Pubblica Luiss Business School

Giappichelli

Filosofia e Questioni Pubbliche
Iscrizione al R.O.C. n. 25223
Registrazione presso il tribunale di Roma n. 290/2000

VIA PO, 21 - TEL. 011-81.53.111
https://www.fqpjournal.com/

ISSN 1591-0660
EISSN 2240-7987

Table of contents

pag.

Contributors III

1. Book Symposia

Symposium on Maeve McKeon's *With Power Comes Responsibility. The Politics of Structural Injustice* (Bloomsbury 2024)

1. Uğur Bulgan and Valentina Gentile, *Structural Injustice in Contemporary Political Theory: An Introduction* 213
2. Vittorio Bufacchi, *Where is the Injustice in Structural Injustice?* 221
3. Mara Marin, *Reflections on Power and Structural Change. Commentary on Maeve McKeown's* With Power Comes Responsibility. The Politics of Structural Injustice 231
4. Rossella De Bernardi, *From Climate Change to Sweatshop Labor: Do "Structural" Injustices Exist, After All?* 241
5. David Owen, *Political Responsibility and the Forms of Solidarity On Maeve McKeown's* With Power Comes Responsibility 251
6. Maeve McKeown, *Response to Critics* 261

2. Special Sections

The Democratic Containment of Fake News and Bad Beliefs

1. Enrico Biale and Gianfranco Pellegrino, *The Democratic Containment of Fake News and Bad Beliefs* 283
2. Neil Levy, *Fake News as Rhetorical Weapon* 289
3. Cathrine Holst, *Democratizing expertise: does the problem of false information change the calculus?* 311

ISSN 1591-0660 | EISSN 2240-7987

pag.

4. Säde Hormio, *The Collective Underpinnings of Bad Beliefs* 337
5. Laura Santi Amantini, *The Unequal Damages of Fake News: Amplifying Epistemic Inequality and Oppression* 365
6. Irene Maria Buso, Margherita Benzi, Marco Novarese and Giacomo Sillari, *Uncertainty and Fake News: An Experimental Study on the Strategic Use of Fake News in Belief Formation* 389
7. Margherita Benzi, Irene Maria Buso, Paolo Chirico, Jacopo Marchetti, Marco Novarese and Giacomo Sillari, *Believe It or Not – An Empirical Study on Fake News Sharing* 407

3. Contemporary Debates in Political Philosophy

1. Olga Lenczewska and Kate Yuan, *Global Poverty, Structural Change, and Role-Ideals* 431
2. Michelle Ciurria, *Strawsonian Responsibility: Three Critiques from the Margins* 459

Contributors

Margherita Benzi is Associate Professor of Logic and Philosophy of Science at the Università del Piemonte Orientale. She published on History and Philosophy of Life Sciences, (with M. Novarese), Argumenta, Topoi, Disputatio, History and Philosophy of Science, Historia Mathematica (and other journals) and is the author of three books: Il ragionamento incerto: probabilità e logica in intelligenza artificiale (1997), Scoprire le cause (2003), Cause singolari (2020).
E-mail address: margherita.benzi@uniupo.it
ORCID: https://orcid.org/0000-0003-4934-6494

Enrico Biale is Assistant Professor of Political Philosophy at the University of Piemonte Orientale. He works on normative democratic theory and theory of social justice. His most recent publications are: Enhancing Democratic Expertise trough Intra-Party Deliberation (with G. Bistagnino), Contellations (2024) and A progressive approach to normative political theorizing (with C. Fumagalli), European Journal of Political Theory (2023).
E-mail address: Enrico.biale@uniupo.it
ORCID: https://orcid.org/0000-0001-7899-0989

Vittorio Bufacchi is Senior Lecturer in the Department of Philosophy at University College Cork, Ireland. He is the author of many books, including *Social Injustice: Essays in Political Philosophy* (Palgrave 2012); *Violence and Social Justice* (Palgrave 2007); *Everything Must Change: Philosophical Lessons from Lockdown* (Manchester 2021); *Why Cicero Matters* (Bloomsbury 2023). He is currently writing a book on the philosophy of human rights.
E-mail address: v.bufacchi@ucc.ie
ORCID: https://orcid.org/0000-0001-7236-117X

Uğur Bulgan is Postdoctoral Research Fellow at LUISS University of Rome, working on the research project, 'Transitional Justice from a Global Perspective: Institutions, Social Norms, and Pluralism'. He is the executive editor of FQP/PPI Journal. His research interests include normative political theory, philosophy of terrorism and counter-terrorism, transitional justice, just war theory, recognition theory, social justice and queer studies. He is the author of "The Just Response to Marital Misrecognition" published in *Biblioteca della Libertà*.
E-mail address: ubulgan@luiss.it
ORCID: https://orcid.org/0000-0002-5556-3869

Irene Buso is Postdoctoral Research Fellow at the University of Bologna. Her main publications include Buso, I. M., Ferrari, L., Güth, W., Lorè, L., & Spadoni, L. (2024). Testing isomorphic invariance across social dilemma games. *Journal of Economic Behavior & Organization*, 223, 1-20; Buso, I.M., & Hey, J. (2021). Why do consumers not switch? *An experimental investigation of a searchand switch model. Theory and Decision*, 91(4), 445-476; Buso, I.M., Di Cagno, D., Ferrari, L., Larocca, V., Lorè, L., Marazzi, F., Panaccione, L., Spadoni, L. (2021). Lab-like findings from online experiments. *Journal of the Economic Science Association*, 7(2), 184-193; Buso, I.M., De Caprariis, S., Di Cagno, D., Ferrari, L., Larocca, V., Marazzi, F., Panaccione, L., Spadoni, L. (2020). The effects of COVID-19 lockdown on fairness and cooperation: Evidence from a lablike experiment. *Economics Letters*, 196, 109577.
E-mail address: irenemaria.buso@unibo.it
ORCID: https://orcid.org/0009-0009-2828-1602

Paolo Chirico: is Researcher at UPO where he teaches Data mining, Statistics, Econometry and Statics methodologies for the impact's evaluation. His main scientific interests include time series analysis, causal inference. His last publications include Chirico, P. (2024) Iterative QML estimation for asymmetric stochastic volatility models. *Statistical Methods & Application*; Bondonio, D. and Chirico, P (2024) "Intertemporal Statistical Matching for Causal Inference in the Context of Multivariate Time-Series Data" in M. Bini, A. Balzanella, L. Masserini, R. Verde (Eds) *Advanced Methods in Statistics, Data Science and Related Applications*, Springer; Novarese, M., Chirico, P. and Di Giovinazzo, V. (2023) Do early freshmen graduate earlier than late ones? Enrolment promptness as an indicator of academic success in In L. Fabbris, S. Mignani, G. Tassinari (eds), *Technology and Data Science for Economic and*

Social Development. Book of Short Papers of the ASA Bologna Conference. Padua, Cleup.
E-mail address: paolo.chirico@uniupo.it
ORCID: https://orcid.org/0000-0003-4229-0440

Michelle Ciurria is a queer, gender-variant, disabled philosopher and professor at Washington University in St. Louis. She completed her PhD at York University in Toronto in 2014 and subsequently held postdoctoral fellowships at Washington University in St. Louis and the University of New South Wales, Sydney. Her research interests include moral responsibility, moral psychology, critical disability theory, Marxist feminism, and decolonial theory. She is the author of *An Intersectional Feminist Theory of Moral Responsibility* (Routledge, 2019) and a regular contributor to BIOPOLITICAL PHILOSOPHY, the leading blog in critical disability theory.
E-mail address: mich.ciurria@gmail.com
ORCID: https://orcid.org/0000-0002-8722-1003

Rossella De Bernardi is Postdoctoral Research Fellow in Political Philosophy at DAFIST, University of Genoa (IT). Previously, she worked at the University of Leeds, King's College London, and the University of Warwick. She holds a PhD (Law, University of York, UK), an MA (Philosophy, University of Pavia, IT), and a BA (Philosophy, University of Genoa, IT). She has interests in political and social philosophy, particularly emotions in social and political life, social injustice, and liberal legitimacy and democratic theory. Her work has appeared in the *Journal of Social Philosophy*, *Journal of Ethics and Social Philosophy*, and *Notizie di POLITEIA*.
E-mail address: rossella.debernardi@edu.unige.it
ORCID: https://orcid.org/0000-0001-6666-3331

Valentina Gentile is Associate Professor in Political Philosophy at LUISS University of Rome. She is the Editor of FQP/PPI Journal. She specializes in normative political philosophy, liberal theory and, especially, the work of John Rawls. Her research focuses on moral stability, pluralism, the principles of reciprocity, toleration and civility and transitional justice. Her work has appeared in several peer-reviewed journals, including *International Theory*, *Journal of Social Philosophy*, *Critical Review of International Social and Political Philosophy* and *Philosophia*. She is author of *From Identity Conflicts to Civil Society* (Luiss University Press, 2013) and co-editor of *Rawls and Reli-*

gion (Columbia University Press, 2015) and *Spaces of Tolerance* (Routledge, 2020). She is currently finalizing a monograph, *Freedom with Religions*.
E-mail address: vgentile@luiss.it
ORCID: https://orcid.org/0000-0002-4080-3444

Cathrine Holst Professor in Philosophy of Science and Democracy, Department of Philosophy, Classics, History of Art and Ideas, University of Oslo. Some recent relevant publications include Cathrine Holst (2024) Worries about philosopher experts, *Res Publica* 30, 47-66, Cathrine Holst & Johan Christensen (2023) The epistemic quality of expert bodies (2023), *Acta Politica* 59, 822-846, Silje Langvatn & Cathrine Holst (2022) Expert accountability: what is it, why is it challenging – and is it what we need? Constellations 31(1), 98-113, Torbjørn Gundersen & Cathrine Holst (2022) Trusted, but not trustworthy? Science advice in an environment of trust, Social Epistemology 36(5), 629-640.
E-mail address: cathrine.holst@ifikk.uio.no
ORCID: https://orcid.org/0000-0002-2231-5826

Säde Hormio Academy Research Fellow and a University Lecturer in Practical Philosophy at the University of Helsinki. Her research focuses on shared and collective responsibility, such as questions about the role of individuals in changing social practices, or what we mean by the responsibility of collective agents. She is also interested in social epistemology, especially institutional ignorance and knowledge. Hormio's publications include *Taking Responsibility for Climate Change* (Palgrave Macmillan, 2024) and "Collective responsibility for climate change" in *Wiley Interdisciplinary Reviews: Climate Change* (2023).
E-mail address: sade.hormio@helsinki.fi
ORCID: https://orcid.org/0000-0002-2231-5826

Olga Lenczewska is Assistant Professor of Philosophy at Florida State University. She specializes in Kant's practical philosophy, contemporary sociopolitical philosophy, and feminist theory. Her most recent publications include the book *Kant on the History and Development of Practical Reason* (Cambridge University Press) and articles such as "Universal Basic Income and Divergent Theories of Gender Justice" (*Hypatia*), "Developing Politically Stable Societies: Kant and Rawls on Moral Maturation" (*Social Theory and Practice*), and "Kant on the Origins of Humanity and Moral Education" (*Journal of the History of Ideas*).
E-mail address: olenczewska@fsu.edu
ORCID: https://orcid.org/0000-0002-7977-8151

Neil Levy is Professor of Philosophy at Macquarie University (Sydney) and a Senior Research Fellow at the Oxford Uehiro Institute. His most recent books are *Philosophy, Bullshit, and Peer Review* (Cambridge University Press, 2023) and *Bad Beliefs: Why They Happen to Good People* (Oxford University Press, 2021).
E-mail address: neil.levy@uehiro.ox.ac.uk
ORCID: https://orcid.org/0000-0002-5507-0300

Maeve McKeown is an Assistant Professor of Political Theory at Campus Fryslân, University of Groningen. She is the author of When Power Comes Responsibility: The Politics of Structural Injustice (Bloomsbury Academic, 2024) and co-editor of What is Structural Injustice? (Oxford University Press, 2024). Her other research interests include reparations for historical injustice and feminism.
E-mail address: m.c.mckeown@rug.nl
ORCID: https://orcid.org/0000-0003-3599-2153

Mara Marin is a political theorist with interests in feminist theory, theories of oppression, domination and structural injustice, as well as the intersections between capitalism, race and gender. Her first book, *Connected by Commitment: Oppression and Our Responsibility to Undermine It*, was published by Oxford University Press. Her articles have been published by *American Political Science Review*, *Contemporary Political Theory*, *Critical Review of International Social and Political Philosophy*, and *Hypatia*, among others. She is currently working on a book manuscript entitled *Structural Agency and Structural Responsibility*. She has a PhD from the University of Chicago and is Assistant Professor at the University of Victoria.
E-mail address: maramarin@uvic.ca
ORCID: https://orcid.org/0000-0003-3959-6887

Jacopo Marchetti is Postdoctoral Fellow at the University of Pisa (Department of Civilization and Forms of Knowledge). He has conducted research in the fields of Political Philosophy and Political Epistemology at the University of Eastern Piedmont, the University of Turin, and the University of Florence. He is the author of two books, *Foucault e Hayek. Tra biopolitica e liberalismo* (2018) and Douglass C. North (2021), and the editor of the book *Il potere della menzogna. Comunicazione e politica nella società digitale* (2024). He has published several articles in international peer-reviewed journals.
E-mail address: jacopo.marchetti@cfs.unipi.it
ORCID: https://orcid.org/0000-0002-2415-0113

Marco Novarese is Associate Professor of Political Economy at UPO. His main publications include: Faralla V., Savadori L., Mittone L., Novarese M., Ardizzone A. (2023), Color and abundance: Influencing children's food choices, *Food Quality and Preference*; Harris M.N., Novarese M., Wilson C.M. (2022), Being in the right place: A natural field experiment on the causes of position effects in individual choice" *Journal of Economic Behavior & Organization*; Benzi M. Novarese M. (2022), Metaphors we Lie by: our 'War' against COVID-19, *History and Philosophy of the Life Sciences*; Faralla, V. & Novarese, M & Di Giovinazzo, V, (2021) "Replication: Framing effects in intertemporal choice with children," *Journal of Economic Psychology*; Faralla, V., Borà, G., Innocenti, A., and Novarese, M. (2020). Promises in group decision making. *Research in Economics*.; Faralla, V; Novarese, M.; Ardizzone, A. (2017) Framing effects in intertemporal choice: A nudge experiment, *Journal of Behavioural and Experimental Economics.*
E-mail address: marco.novarese@uniupo.it
ORCID: https://orcid.org/0000-0001-5622-5549

David Owen is Professor of Social and Political Philosophy at the University of Southampton and currently SSS Visiting Professor at IAD, Princeton. His recent work focuses on the ethics of displacement and on the concept of vindication in ethics and political philosophy.
E-mail address: dowen@ias.edu
ORCID: https://orcid.org/0000-0001-8016-2947

Gianfranco Pellegrino is Full Professor in Political Philosophy at Luiss Guido Carli. He has recently edited the Springer's Handbook of the *Philosophy of Climate Change* (2023) and *Canned heat: ethics and politics of global climate change* (Routledge, 2014). His work has appeared in several journals including *Ethical Theory and Moral Practice* and *Philosophical Writings*.
E-mail address: gpellegrino@luiss.it
ORCID: https://orcid.org/0000-0002-8029-3936

Laura Santi Amantini is Postdoctoral Research Fellow at UPO. She received her PhD in Philosophy from the University of Genoa. She held visiting positions at the University of Southampton, Bristol and Oxford. Her previous work has appeared in the *Journal of Social Philosophy*, *Res Publica*, *Global Justice: Theory Practice Rhetoric*, *Biblioteca della Libertà*, and *La Società degli Individui*. She also contributed a chapter for the volume *The Political*

Philosophy of Internal Displacement (Oxford University Press 2024, ed. by J. Draper and D. Owen).
E-mail address: laura.santi_amantini@uniupo.it
ORCID: https://orcid.org/0000-0001-5604-9389

Giacomo Sillari is Associate Professor of Philosophy of Science at Luiss Guido Carli University. He holds a Ph.D. from Carnegie Mellon University and has held positions at the University of Pennsylvania, Scuola Normale Superiore, and various visiting professorships. His research spans philosophy, logic, game theory, and behavioral sciences, with applications to social sciences, economics, and public policy. Prof. Sillari has published in top journals including *Synthese*, *Mind and Society*, *J of Phil Logic*, *Philosophical Studies*, contributed to edited volumes published by, among others, Oxford University Press, and served on editorial boards.
E-mail address: gsillari@luiss.it
ORCID: https://orcid.org/0000-0002-3243-1206

Kate Yuan is a PhD student in philosophy at Yale University. Her research interests lie in social and political philosophy and applied ethics, with a focus on issues at the intersection of global justice, feminist philosophy, and public policy. Her most recent publication, "Global Justice: From Institutional to Individual Principles," appears in *Social Theory and Practice*.
E-mail address: kate.yuan@yale.edu
ORCID: https://orcid.org/0009-0004-0923-7968

Book Symposia

Symposium on Maeve McKeon's *With Power Comes Responsibility. The Politics of Structural Injustice* (Bloomsbury 2024)

DOI 10.17473/2240-7987-2024-2-1

Structural Injustice in Contemporary Political Theory: An Introduction*

Uğur Bulgan** and Valentina Gentile***

Abstract

McKeown's book, *With Power Comes Responsibility: The Politics of Structural Injustice* (Bloomsbury 2024) revisits Iris Marion Young's theory of structural injustice, incorporating critical realism and adding a focus on power dynamics with the aim of clarifying the contours of political responsibility when systemic inequalities are at stake. In this introduction, we first reconsider Young's original idea of structural injustice in light of some important critiques raised in contemporary political theory literature. We then present key aspects of McKeown's reformulation of this idea.

Summary: I. Structural Injustice and its Critics. – II. McKeown's contribution to the debate. – Works Cited.

We live in an unjust world. Some of these injustices result directly from the actions of agents. If one person is robbed on the street, or if many are harmed by a defective drug, or if a civilian population is disproportionately

* The idea of structural injustice has recently attracted renewed interest in contemporary political theory. For the purposes of this introduction, we thought it appropriate to draw on some of this debate to give readers a sense of where it stands. Maeve McKeown's *With Power Comes Responsibility: The Politics of Structural Injustice* (Bloomsbury 2024), the subject of this symposium, is an important contribution to this literature. We are grateful to Maeve McKeown and all the contributors to this symposium for their thoughtful reflections on this book. The two authors contributed equally to the writing of the introduction.

** ORCID: 0000-0002-5556-3869.

*** ORCID: 0000-0002-4080-3444.

targeted in a war, the perpetrator(s) of the injustice can be identified and possibly punished. In these cases, the very existence of an injustice is linked to one or more wrongdoers. Yet, consider the case of Jane.[1] Jane is a single mother of three children, who has just been fired from her job as a part-time store clerk. Having lost her job, Jane can no longer afford to rent the apartment where she used to live with her children and is forced to move to a more suburban area. Jane finally manages to find a small apartment in a public housing project. However, after a few weeks, there is an earthquake in the area. The public housing unit where Jane lives with her children is declared uninhabitable. Jane thus finds herself jobless and homeless, and the effects of an external event such as an earthquake have a disproportionate impact on her living conditions. Jane is clearly facing some form of injustice, and yet in this case it is difficult to make a connection between the injustice she is experiencing and one or more wrongdoers. This kind of injustice, as Iris Marion Young has famously argued, actually takes a structural form. For Young (2011), the injustice people like Jane face results from the aggregate actions of multiple agents within existing social structures.[2] The fact that injustice is structural does not, however, have any bearing on our interest in the concept of responsibility. Yet, what kind of responsibility? According to Young and in contrast with standard legal and moral models of responsibility, we should embrace a social connection account of responsibility. This model links existing structural forms of injustice to a forward-looking, prescriptive ideal of political responsibility. Recently, a great deal of academic work has emerged which focuses on the notion of structural injustice and the associated view of political responsibility. In this context, Maeve McKeown's *With Power Comes Responsibility: The Politics of Structural Injustice* (Bloomsbury 2024) not only offers a unique contribution to a deeper understanding of structural injustice but also helps to locate this concept and its theoretical implications within today's political theory.

[1] The case of Jane, elaborated in the following discussion, draws inspiration from Iris Marion Young's seminal Sandy's case, which has become the paradigmatic example of structural injustice. To bridge Young's foundational insights with contemporary critical perspectives on structural injustice, and especially Estlund's recent contribution (Estlund 2024), we have incorporated a natural disaster into the scenario.

[2] For Young's elaboration on structural injustice, social connection model and political responsibility *see also* Young, 1990; 2003; 2004; 2006a; 2006b.

I. Structural Injustice and its Critics

It is undeniable that the concept of structural injustice has a strong explanatory power. We could say that it somehow captures the *zeitgeist* of our societies. It is therefore no surprise that in recent years, an increasing number of political theorists have recovered this notion expanding its application to various cases, such as colonialism (Lu 2011; 2017; 2023; Ypi, 2017), gender inequality (Nuti 2019; Parekh 2011), and climate change (Godoy 2017). As is often the case with powerful concepts, however, such academic attention is not limited to that part of the literature which takes up and defends the original idea of structural injustice. Indeed, this literature has provided important insights for a critical reading of this approach to both injustice and responsibility. Andrea Sangiovanni, for example, has recently argued that moral responsibility should be attributed to individuals as long as these actions collectively produce or reproduce unjust structures, and concludes that the responsibility for structural injustice should be both backward and forward-looking (Sangiovanni 2018). It is again Young's distinction between backward and forward-looking responsibility that is at the core of Robert Goodin's and Christian Barry's recent contribution to this debate. While sympathizing with Young's general project, the two authors highlight that this distinction results motivationally ineffective as, in their own reasoning, why should people feel compelled to act with responsibility for the future, knowing that if they fail to fulfil it, they will be absolved of any subsequent guilt that looks to the past (Goodin & Barry 2021, 340)? They also point to an internal incoherence in this model, since the same reasons for not attributing backward-looking responsibility would also apply to forward-looking responsibility (*Ibid.*). More recently, David Estlund (2024) has critically examined the philosophical and moral underpinnings of structural injustice. He compares structural harms to natural disasters, questioning whether social structures inherently warrant grievance attitudes in the absence of wrongdoing. He concludes that while structural injustice remains an essential concept, its classification as "wrong" versus "bad" requires careful handling, especially with regard to attitudes of grievance, resentment and moral indignation in the absence of identifiable wrongdoers.

II. McKeown's contribution to the debate

McKeown's *With Power Comes Responsibility,* which is the basis of this symposium, is also important in light of these critiques as far as it offers a

comprehensive attempt to critically re-examine structural injustice, by confronting it with today's most pressing political challenges, while remaining committed to Young's original project. The book introduces two important amendments to the original theory: first, it is based on a critical realist ontology that aims to deepen the conceptual foundations of Young's original paradigm; and second, it introduces the idea of power as linked to political responsibility. In other words, it offers new ontological, analytical and normative tools for approaching structural injustice. As editors of this symposium, we would like to point out that this book is one of the most sophisticated, yet still critical, recent tributes to Young's project.

By situating the structural framework within Archer's (1995) critical realist approach, McKeown aims to overcome the weaknesses of Gidden's structuration theory (Giddens 1979; 1984) which, she claims, is ontologically inconsistent in its approach to the relationship between structure and agency. For her, critical realism explicitly separates structure and agency and allows for a more nuanced understanding of their interplay. McKeown uses this more nuanced approach to analyze how power operates within social structures, arguing that powerful agents have a greater capacity to influence and shape structures. Thus, for the author, a critical theory approach to structural injustice must incorporate a thorough analysis of power. After accurately dissecting five dimensions and three forms of power, she reveals how power dynamics contribute to the creation, maintenance and perpetuation of systemic inequalities. Descending the ladder of abstraction, she examines the power of multinational corporations in the global political economy, particularly in the garment industry.

McKeown's analysis of power helps to differentiate types of structural injustice by showing how different levels of power influence agents' actions. Such analysis moves beyond envisioning structural injustice as an unintended consequence of individual actions. Structural injustice is in fact "deliberate" when powerful agents deliberately perpetuate unjust structures in order to exploit disadvantaged groups for their own benefit (McKeown 2024, 44). It is also "avoidable" when powerful agents do not deliberately perpetrate the injustice but fail to remedy it, despite having the capacity to do so (McKeown 2024, 43). And, structural injustice is "pure" when injustice arises entirely from social processes with no single powerful agent being able to remedy it (McKeown 2024, 41). To understand how such an analysis of power should inform the idea of responsibility, it is important to refer to McKeown's (2024, 36) scheme for assigning different responsibilities to powerful agents and ordinary individuals. She critically examines conceptions of moral and political

responsibility and proposes a "political conception of political responsibility" (McKeown 2024, 143) that requires ordinary citizens to develop the capacity for political solidarity. She argues that ordinary citizens are not morally responsible for structural injustice, but that corporations bear moral responsibility for historical structural injustice (McKeown 2024, 201).

Each commentary included in this symposium brings a unique perspective that contributes to a deeper understanding of McKeown's arguments, highlighting the book's novel approach to structural injustice while critically engaging with it.

Vittorio Bufacchi praises McKeown's approach to structural injustice but stresses that the rich discussion of structural harm and violence within the Marxist tradition is overlooked. Bufacchi notes that McKeown's tripartite typology of structural injustice relies heavily on "the levels of intentionality of social actors vis-à-vis the structures in which they operate." However, he also believes that McKeown's account lacks an analysis of relevant concepts from the philosophy of action, such as intentionality and foreseeability. At the same time, he is concerned about McKeown's reliance on power, arguing that without deeper analysis, her typology may remove the structural nature of structural injustice. His final concern relates directly to McKeown's interpretation of power, suggesting the need to distinguish powerlessness from disempowerment.

Mara Marin underlines two major contributions of McKeown's book: first, it provides conceptual clarity by separating moral from political responsibility; second, it helps to identify levels of agents' responsibility based on their power. However, she also raises important criticisms. First, she is concerned about McKeown's reliance on powerful agents as the primary drivers of change. Indeed, structural change may require the harmonized influence of less powerful agents. She also suggests that McKeown's typology of structural injustice risks diminishing the importance of a structural lens by focusing on agent-specific accountability, potentially conflating structural and agentive forms of injustice. Marin concludes that while McKeown's framework may redefine understandings of structural injustice by emphasizing power, it also highlights the need for further exploration of how various forms of agency and structural positioning influence responsibility and justice.

Rossella De Bernardi's commentary explores McKeown's integration of power dynamics into Young's original paradigm. She agrees with McKeown that corporations have unique agency within structural systems, and hence special responsibilities. However, she questions the very idea of "deliberate" structural injustice. She wonders whether the labelling of certain injustices as

"structural" is still valid when intentional agents can cause or change them. Using examples like sweatshop labor and global poverty, De Bernardi discusses the consistency of assigning moral responsibility to powerful actors in structurally unjust contexts. She suggests that distinguishing between unjust harms and structural causation remains problematic, especially when corporations might be blamed for sustaining broad systemic injustices beyond direct harms.

David Owen's contribution focuses on political responsibility and political solidarity. He emphasizes the gap between the theoretical framework and how the examples are treated. Owen delves into two key areas in which he believes McKeown's examples reveal implicit but significant philosophical commitments: decision-making power in tackling injustice and the nature of solidarity. First, he argues that giving those most affected by structural injustice a central role in decision-making is not just about respecting the insights of victims but is a matter of justice. McKeown's discussion of the anti-sweatshop movement shows how the perspectives of sweatshop workers should shape activism to avoid harm, such as unwanted job losses from boycotts. Second, Owen examines McKeown's account of political solidarity while distinguishing between symmetrical (within a marginalized group) and asymmetrical forms (privileged individuals supporting marginalized groups). The example of the feminist movement shows a symmetrical, in-group solidarity whilst the collaboration between United Students Against Sweatshops and sweatshop workers demonstrates asymmetrical out-group solidarity. Owen argues that both types of solidarity are essential, especially given intersectional differences within groups. The symposium concludes with a rejoinder by Maeve McKeown.

Works Cited

Archer, Margaret. 1995. *Realist Social Theory: The Morphogenetic Approach*. Cambridge: Cambridge University Press.

Estlund, David. 2024. "What's Unjust about Structural Injustice?". *Ethics* 134 (3).

Giddens, Anthony. 1979. *Central Problems in Social Theory: Action, Structure and Contradiction in Social Analysis*. London: Macmillan Press.

Giddens, Anthony. 1984. *The Constitution of Society: Outline of the Theory of Structuration*. Cambridge: Polity Press.

Godoy, Eric. 2017. "What's the Harm in Climate Change?". *Ethics, Policy and Environment* 20 (1): 103-117.

Goodin, Robert, and Christian Barry. 2021. "Responsibility for structural injustice: A third thought." *Politics, Philosophy & Economics* 20 (4): 339-356.

Lu, Catherine. 2011. “Colonialism as structural injustice: Historical responsibility and contemporary redress.” *The Journal of Political Philosophy* 19 (3): 261-281.
Lu, Catherine. 2017. *Justice and reconciliation in world politics*. Cambridge: Cambridge University Press.
Lu, Catherine. 2023. “Progress, decolonization and global justice: a tragic view.” *International Affairs* 99 (1): 141–159.
McKeown, Maeve. 2024. *With Power Comes Responsibility: The Politics of Structural Injustice*. London: Bloomsbury.
Nuti, Alasia. 2019. *Injustice and the Reproduction of History: Structural Inequalities, Gender and Redress*. Cambridge: Cambridge University Press.
Parekh, Serena. 2011. “Getting to the root of gender inequality: structural injustice and political responsibility”. *Hypatia* 26 (4): 672-689.
Sangiovanni, Andrea. 2018. “Structural Injustice and Individual Responsibility.” *Journal of Social Philosophy* 49 (3): 461–483.
Young, Iris Marion. 1990. *Justice and the Politics of Difference*. New Jersey: Princeton University Press.
Young, Iris Marion. 2003. “From guilt to solidarity: Sweatshops and political responsibility.” *Dissent* 50 (2).
Young, Iris Marion. 2004. “Responsibility and global labor justice.” *The Journal of Political Philosophy* 12 (4): 365–388.
Young, Iris Marion. 2006a. “Katrina: Too much blame, not enough responsibility.” *Dissent* 53 (1): 41-46.
Young, Iris Marion. 2006b. “Responsibility and global justice: A social connection model.” *Social Philosophy and Policy* 23(1): 102-130.
Young, Iris Marion. 2011. *Responsibility for justice*. Oxford: Oxford University Press.
Ypi, Lea. 2017. “Structural Injustice and the Place of Attachment.” *Journal of Practical Ethics* 5 (1).

DOI 10.17473/2240-7987-2024-2-2

Where is the Injustice in Structural Injustice?

Vittorio Bufacchi*

Abstract

This paper will offer a conceptual analysis and critique of Maeve McKeown's account of structural injustice. McKeown's main thesis is that structural injustice ought to be approached via critical theory, and that a critical theory of structural injustice should incorporate power. While I agree with McKeown's general approach, I argue that two aspects of this work remain opaque and in need of further analysis. First, while McKeown does a lot of important work on the 'structural' component of structural injustice, what constitutes an 'injustice' remains undefined, being taken for granted. More specifically, the concept of intentionality, which is crucial to McKeown's account of 'deliberate structural injustice', remains under analysed. Secondly, while McKeown rightly puts power at the centre of structural injustice, there seems to be an assumption that power and domination are cut from the same cloth. I will challenge this assumption, suggesting an alternative way of integrating power within the sphere of structural injustice.

Summary: Introduction. – The Main Thesis. – I. Structural Injustice: Does Intentionality Matter? – II. Where is the Injustice? – Conclusion. – Works Cited.

Introduction

Let me be as clear as I can be from the outset. This is a superb work of scholarship that ought to be read by anyone seriously interested in matters of social justice, and structural injustice in particular. Maeve McKeown (2024) has written the book that many of us working in the field have been waiting for at least since 1990, the year when Iris Marion Young's *Justice and the Politics of Difference* was published. McKeown ought to be congratulated and praised for writing this book. If there is any justice in the world of academia,

* ORCID: 0000-0001-7236-117X.

this book will be recognised as a major achievement. Nothing I say in the following pages takes away from this verdict.

The Main Thesis

McKeown summarizes her book in one sentence: "My main thesis is that a critical theory of structural injustice should incorporate *power*" (2). There are two key terms here: 'critical theory' and 'power'. By 'critical theory' McKeown is referring to the work by Margaret Archer, in particular her 1995 book *Realist Social Theory*. McKeown adopts Archer's critical theory to correct certain aspects of Young's work. The other key term is 'power', arguably the real protagonist in McKeown's book, as McKeown indicates: "the aim of this book is to integrate power into structural injustice theory" (15). Regarding power, McKeown leans heavily on Thomas Wartenberg's analysis taken from his book *The Forms of Power*.

With the help of these two authors, Archer and Wartenberg, and their respective accounts of critical theory and power, McKeown constructs an elegant, detailed analysis of structural injustice. According to McKeown, there isn't just one but three types of structural injustice. In the first half of the book McKeown offers her original account of structural injustice. McKeown's sophisticated philosophical approach to structural injustice is accompanied by many real-life examples, making her work accessible, relevant, and politically poignant: sweatshops; poverty; climate change; corporate abuse and exploitation. In the second half of the book McKeown theorizes responsibility for structural injustice. In what follows I will focus exclusively on the first half of McKeown's book.

I. Structural Injustice: Does Intentionality Matter?

McKeown is inspired by the work of Iris Marion Young (1990). This is not surprising, since every scholar who writes about structural injustice today inexorably lists Young as their chief point of reference, and McKeown is not an exception to this rule. While this is perfectly understandable, it is also slightly disappointing. The idea of structural injustice predates Young, and no one should think that Young invented the concept or coined the term. Anyone familiar with the work of Marx knows that structural injustice has a long history.

Even if we disregard Marx's observations that under capitalism *everyone* is

alienated, both capitalists and proletarians, hence the capitalist mode of production forms the structure within which injustice is articulate, and we ignore the works of other Marxists in the 20th century, starting with Rosa Luxemburg to Antonio Gramsci, there is an important literature in the late 60s and early 70s on structural injustice that seems to have gone under the radar. I'm referring to Johan Galtung and Newton Garver *in primis*. Although they write about 'structural violence' and 'institutional violence', they also have a very broad conception of violence which overlaps with injustice, in fact Galtung often uses the terms 'structural violence' and 'social injustice' interchangeably.

Johan Galtung (2009, 83) famously distinguished between 'direct violence', where the instigator of an act of violence can be traced to a person or persons, and 'structural violence', where violence occurs but cannot be traced back to any person who directly harms another person: "there may not be any person who directly harms another person in the structure. The violence is built into the structure and shows up as unequal power and consequently as unequal life chances". Working in a similar vein, Newton Garver (2009) distinguishes between four types of violence, creating a 2 X 2 matrix: violence can be personal or institutional, and it can be overt or covert.

I have always wondered why Young failed to engage with the works of Marx, Galtung or Garver in a meaningful way. Be that as it may, Young must be given credit for redirecting the vast literature on social justice to questions about social injustice, and more specifically on the structural basis of injustice. One of the things that is slightly disappointing about the current literature on structural injustice is that the reverence towards Young can reach levels akin to a religious cult. Her views are often endorsed and reiterated unapologetically without any hints of criticism. Thankfully this is not the case with McKeown. While McKeown is very approving of Young, it is refreshing to see that she is also not afraid to be critical of Young. In suggesting shortcomings in Young's work, McKeown is putting forward a constructive criticism with the aim of improving and strengthening Young's theory. This is exactly as it should be, and it's duly welcomed.

One of the major virtues of McKeown's book is to explore in greater detail the concept of 'structure', something that is often assumed to be self-explanatory in the literature on structural injustice. It isn't of course. As McKeown rightly points out, the interaction between structure and individual action remains under-analysed and in need of further clarification.

McKeown distinguishes between three types of structural injustice: pure, avoidable, and deliberate. Of these, deliberate structural injustice is perhaps the most interesting, but also the most problematic. 'Pure' structural injustice

captures the idea that social actors cannot escape from reproducing the injustice, therefore the consequences of their actions are unintended. 'Avoidable' structural injustice posits that not all agents are objectively constrained by the structures, allowing for the fact that some agents have the power to change the unjust structures but fail to do so, therefore the outcomes may be unintended but foreseeable and avoidable. 'Deliberate' structural injustice goes one step further. McKeown defines deliberate structural injustice as follows: "structural injustice is 'deliberate' when the unjust outcomes are intended because powerful agents benefit from it so they deliberately perpetuate it, and these agents have the capacity to remedy it but they fail to do so" (45).

The difference between pure, avoidable and deliberate structural injustice seems to come down to the levels of intentionality on the part of the social actors vis-à-vis the structures in which they operate: 'Pure' is unintentional; 'Avoidable' is unintentional but foreseeable; 'Deliberate' is intentional.

This typology is valuable, and McKeown ought to be congratulated for introducing these distinctions. But given the centrality of intentionality and unintentionality to her analysis, it is surprising that McKeown fails to expand on the concept of intentionality. We are in the realm of philosophy of action here, not political philosophy, nevertheless political philosophers disregard this body of literature at their own peril. There is also an interesting literature on intentionality in relation to acts of violence, in fact on the question of intentionality there are instructive parallels between structural injustice and violence.

In my book *Violence and Social Justice,* I argue that while most definitions of violence assume intentionality, in the sense that the action by X was undertaken with the deliberate aim of causing harm to Y, it is also possible and desirable to define violence as the outcome of an unintentional act (Bufacchi 2007).

Consider these three cases:

1. X performs action A with the intended aim of doing good to X but action A has the foreseen but unintended consequence of also harming Z.
2. X performs action A with the intended aim of doing harm to X but action A has the foreseen but unintended consequence of also harming Z.
3. X performs action A with the intended aim of doing *some* harm to X but action A has the foreseen but unintended consequence of doing *much greater* harm to X.

The concepts of intentionality and foreseeability are crucial here, furthermore these distinctions apply to structural injustice as much as to violence. In

philosophy of action the orthodox view (often associated with Donald Davidson) is that behind the idea of actions is the notion of agency, and intentionality implies reason: to say that person P acted intentionally is to say that P performed action A because P had a reason to A. But this view is not shared by everyone. Michael Bratman (1987) argues that apart from cases where action A is intentional whenever the agent performing the action intends to do A, there are also cases when what we do intentionally does not fully match with what we intended to do. Bratman is suggesting that we need to distinguish between what we do intentionally, and what is intended. When an intended action has foreseen but undesired side effects, the agent will have brought about the effects intentionally, but not to have intended to bring this about. [1] Given the central role of intentionality in McKeown's theoretical framework, it is surprising that this concept does not get the detailed analysis it deserves.

Leaving the question of intentionality aside, there is another issue regarding McKeown's concept of deliberate structural injustice worth considering. When political philosophers began to investigate the structural nature of injustice, one of the distinctive features of this concept was that no single person is responsible for the injustice. Thus, while there are many cases of single individuals deliberately acting in ways that cause or perpetuate an injustice on other individuals, in which case the injustice is strictly interpersonal, to think of injustice structurally was inviting us to switch our attention to the social, political, economic, and cultural context within which individuals operate. To put it bluntly, while interpersonal injustice has a distinct Kantian flavour (what make injustice wrong can be attributed to the reasons and the free will of its perpetrators), structural injustice has a manifest Marxist flavour (what make injustice wrong can be attributed to the context in which agents operate, in Marx's case the capitalist mode of production). [2]

This idea of structural injustice beyond intentionality is captured, for example, by Newton's Garver's concept of institutional violence. Garver (2009, 180) tells us that "the institutional form of quiet violence operates when people are deprived of choices in a systematic way by the very manner in which transactions normally take place, without any individual act being violent in itself or any individual decision being responsible for the system". But now

[1] On intentionality and foreseeability see also Gilbert Harman (1986) and John Harris (1980).

[2] Of course, this is assuming that Marxism is an ethical theory, and that a Marxist theory of social justice is not an oxymoron. We will not go there here, but for the records I'm with Norman Geras (1989) on this question.

McKeown's idea of deliberate structural injustice wants to reintroduce intentionality within the scope of structural injustice. While there are merits to this, there is also the risk that the most distinctive element of this injustice is stripped away. In other words, once the intentionality of the agent is part of the structural injustice equation, we are back in the realm of personal injustice. It is true that interpersonal injustice can be direct or indirect, and of course McKeown is trying to capture the way agents intentionally use certain structures to inflict an injustice, but sometimes I fear that the most radical dimension of structural injustice is compromised once personal intentionality is once again back in the fray.

II. Where is the Injustice?

As we have seen, much of the focus of McKeown's book is on the concept of structure in structural injustice. Here McKeown makes extremely useful observations that break new grounds in the literature on social injustice. Of course, the concept of 'structural injustice' is not only about the 'structure', it is also (arguably primarily) about 'injustice'. It is therefore surprising that while McKeown has a lot to say about structures, in some ways she has less to say about injustice. At the start of Chapter 2 McKeown writes: "Structural injustice, broadly speaking, is the fallout of social-structural processes that render groups of people vulnerable to domination or oppression" (19). I don't disagree with this, but 'vulnerability', 'domination' and 'oppression' are complex terms that need to be defined and analysed: vulnerable to what? Dominated by whom, and how? And who decides whether a group is oppressed or not? Also, why groups and not individuals? And finally, is injustice about 'rendering' groups vulnerable, or taking advantage of their vulnerability? I couldn't find satisfactory answers to these questions.

Later in the book McKeown expands on these issues, even though often by pointing the reader to the work of Young. Thus, in wording that echo Young's, McKeown says that "domination prevents individuals from determining how they will live their lives. Oppression prevents individuals from developing their unique capacities and fulfilling their personal potential" (p.36). I was not persuaded by Young's take on oppression, and I'm not convinced by McKeown's claims regarding domination either. [3] Defining oppres-

[3] For my critique of Young, see Bufacchi (2012).

sion in terms of personal potential not being fulfilled is highly problematic. I have never met a person who has fulfilled their personal potential, and if an injustice occurs whenever someone does not fulfil their personal potential, then we might as well just give up: injustice is ubiquitous and unstoppable, and will never be overcome.

Accidentally, the emphasis on one's personal potential being unfulfilled is also central to the views on social injustice and structural violence advanced by Galtung and his many followers. Thus, Galtung (1969, 168) writes that "Violence is present when human beings are being influenced so that their actual somatic and mental realizations are below their potential realizations". [4] Galtung's account of structural violence has been criticised on this issue, and rightly so, for being much too broad. The same criticism applies to Young on oppression and to McKeown on domination.

To give credit where credit is due, McKeown's account of structural injustice also has an appealing and innovative twist, for McKeown puts a great deal of emphasis on the concept of power, which distinguishes her work from Young's. I welcome this move, this is arguably the strongest aspects of McKeown's book, even though I have some reservations about McKeown's analysis of power.

First, following Wartenberg, McKeown seems to think of power in binary terms: there are the powerful, and the powerless. Once again, we find the fingerprints of Young on this dichotomy. Young (1990, 52) tells us that one of the five faces of the five faces of oppression is powerlessness: "The powerless are those who lack authority or power even in this mediated sense, those over whom power is exercised without their exercising it; the powerless are situated so that they must take orders and rarely have the right to give them". There is nothing wrong with this definition of powerlessness, but I also think that oppression is not so much about powerlessness, instead it is about disempowerment. The difference is subtle but important. While powerlessness denotes a state of affairs, or a state of being, disempowerment is an active process which delineates how power is taken away from someone. Disempowerment is dynamic, while powerlessness is static. I wish there was room for disempowerment and not only powerlessness in McKeown's analysis. [5]

Leaving aside the distinction between powerlessness and disempowerment, there is another issue regarding McKeown's account of power that I don't find

[4] See also Jamil Salmi (1993).

[5] I discuss social injustice in terms of disempowerment in Bufacchi (2012), pp. 14-15.

totally persuasive. Once again following on Wartenberg's footsteps, McKeown stresses the affinity between power and domination. In fact, McKeown goes as far as to adopt Wartenberg's definition of domination: "Domination refers to a relationship between social groups in which 'power is exercised by the dominating social agent repeatedly, systematically, and to the detriment of the dominated agent'" (p.39). While one can't deny that at one level power and domination are closely related, the two concepts should be distinguished. The move from power to domination is too quick, and potentially problematic on three accounts.

First, power is a dispositional concept. Peter Morriss (1987, 49), who wrote arguably still the best philosophical analysis of this concept, says that power is "a sort of ability: the basic idea is that your powers are capacities to do things when you choose". To think of power as an ability is not the same as saying that to have an ability is to have power. As Brian Barry (1988) rightly points out, 'power' and 'ability' are not interchangeable, and not all ability is power. For Barry (1988, 341): "there is more to (social power) than being able to do what you choose to do. Your having power entails that you have the ability to overcome resistance or opposition and by this means achieve an outcome different from the one that would have occurred in the absence of your intervention". The point about power being an ability is that the concept of power itself is normatively neutral, whereas domination has clear negative connotations.

Second, McKeown says that "the dominant agent does not need to issue a threat to coerce the subordinate, rather the dominant agent's ability to harm the subordinate is a structural feature of their relationship" (p. 39). This is true, but in that case, as Brian Barry (1980, 183) would say, the dominant agent is merely lucky, not powerful: "Must the power of a group be conceptualized as the sum of the power of each of the individual members of the group, or could a group be powerful whilst each of its members is individually powerless?". The difference between luck and power is analytically significant, in part because outcomes do not necessarily reveal power. Keith Dowding (2019, 47) offers a very good example to highlight the difference between outcomes and power: "If the conservative bloc on the Supreme Court have a clear majority, say six to three, then as a group the conservatives are clearly powerful. On all 'ideological' decisions they can get what they want. However each conservative Judge, as an individual, has exactly the same power as every other, and as each of the liberal Justices: viz. one vote each". This example also raises questions about the rudimentary relationship between domination and power.

Third, and perhaps more importantly, McKeown's account of power doesn't really tell us much about how power is used, or what forms it takes.

The point that McKeown wants to make is that some agents intentionally use or interact with existing structures in order to dominate others. I don't have a problem with this, in fact I agree with it. But we need to understand the ways in which agents use their power, or in other words we need to understand what is distinctive about their ability. Here it is important to distinguish between 'outcome power' (power *to*: the power to bring about outcomes) and 'social power' (power *over*: a social relation between at least two people). If we focus on social power, it is necessary to understand precisely what form social power takes.

Keith Dowding (2019, 48) has an interesting take on this: "Social power is the ability of an actor deliberately to change the incentive structure of another actor or actors to bring about, or help bring about outcomes". I think this is correct. The importance of Dowding's analysis is to remind us that too often we attribute power where it isn't there. In the literature this is referred to as the political power fallacy. Dowding (2019, 88) is right when he reminds us that the fact that actor A is powerless to bring about some outcome *x* does *not* imply that there is another actor B who is powerful enough to stop her. "The inference is false. Even if there is an actor B who is powerful enough to stop them bringing about *x*, the fact that A cannot do so is not sufficient to demonstrate that B is to blame". What Dowding is getting to here is that often what makes a certain group powerless is the fact that this group is unable to mobilize itself, that they struggle to overcome their own collective action problem. If certain individuals or another group benefits from this, they are simply lucky, not necessarily powerful. So how does power manifest itself? Dowding's answer is ingenious: an agent is powerful when they deliberately change the incentive structure of another agent so that they fail to mobilize, and remain victim of their own collective action problem. I find Keith Dowding's analysis of power very persuasive, and I believe it could be integrated within McKeown's theoretical framework to great effect.

Conclusion

McKeown has written an extremely important book, that one day may replace Young's work as the standard reference point on structural injustice. In this article I have raised some minor questions about the role of intentionality in the definition of structural injustice, and the ways in which power operates. But nothing I have written takes away from the immense value of McKeown's book, denoting a very significant contribution to contemporary political philosophy.

Works Cited

Margaret Archer, Margaret. 1995. *Realist Social Theory*. Cambridge: Cambridge University Press.

Barry, Brian. 1980. "Is It Better To Be Powerful or Lucky?", Parts 1 and 2. *Political Studies* 28: 183-94 and 338-52.

Barry, Brian. 1988. "The Uses of 'Power'". *Government and Opposition* 23 (3): 340–53.

Bratman, Michael. 1987. *Intentions, Plans, and Practical Reason*. Cambridge, Mass: Harvard University Press.

Bufacchi, Vittorio. 2007. *Violence and Social Justice*. London: Palgrave.

Bufacchi, Vittorio. 2012. *Social Injustice: Essays in Political Philosophy*. London: Palgrave.

Dowding, Keith. 2019. *Rational Choice and Political Power*. 2nd Edition. Bristol: Bristol University Press.

Geras, Norman. 1989. "The Controversy About Marx and Justice", in *Marxist Theory*, edited by A. Callinicos. Oxford: Oxford University Press.

Galtung, Johan. 1969; 2009. "Violence, Peace and Peace Research", in *Violence: A Philosophical Anthology*, edited by V. Bufacchi. London: Palgrave.

Garver, Newton. 1968; 2009. "What Violence Is", in *Violence: A Philosophical Anthology*, edited by V.Bufacchi. London: Palgrave.

Harman, Gilbert. 1986. *Change in View*. Cambridge, Mass.: Harvard University Press.

Harris, John. 1980. *Violence and Responsibility*. London: Routledge.

McKeown, Maeve. 2024. *With Power Comes Responsibility: The Politics of Structural Injustice*. London: Bloomsbury.

Morriss, Peter. 1987. *Power: A Philosophical Analysis*. 2nd Edition. Manchester: Manchester University Press.

Salmi, Jamil. 1993. *Violence and Democratic Society*. London: Zed Books.

Wartenberg, Thomas. 1990. *The Forms of Power: From Domination to Transformation*. Philadelphia: Temple University Press.

Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University Press.

DOI 10.17473/2240-7987-2024-2-3

Reflections on Power and Structural Change

Commentary on Maeve McKeown's *With Power Comes Responsibility. The Politics of Structural Injustice*

Mara Marin*

Abstract

Maeve McKeown's *With Power Comes Responsibility* (WPCR) convincingly argues that discussions of structural injustice and responsibility for it should integrate discussions of power relations. Powerful agents have different responsibilities than "ordinary individuals" because they have access to more resources and have more "elbow room" to make changes. However, WPCR focuses on one form of power – the power agents have in virtue of their structural position – and assumes that this form of power always translates in power to change structures. This is a mistake because the structurally privileged are not necessarily better able to change structures. Men, for example, are not necessarily better able than women to change sexist structures. All occupants of structural positions arguably control resources. In some cases, the subordinates have a monopoly over important resources. In a society that assigns caregiving responsibility exclusively to women, for example, women have a monopoly over caregiving skills and knowledge. This control of resources gives the subordinate – understood as a collective, not individual agent – the power to change the structure in virtue of which they have those resources by acting in ways that do not conform to their mandated use in the structure.

Maeve McKeown's *With Power Comes Responsibility* (*WPCR*) is a major contribution to the structural injustice literature. It contributes to at least two lines of inquiry, each of which would have qualified it as an impressive work.

First, *WPCR* clarifies and deepens our understanding of Iris Young's view of structural injustice, highlighting rather than glossing over the ambiguities,

* ORCID: 0000-0003-3959-6887.

tensions and theoretical lacunae in the posthumously published in 2011 (and not finished in 2006, at the time of Young's death) *Responsibility for Justice*, the work responsible for coining the notion of "structural injustice" and introducing it to mainstream discussions in political theory. McKeown is a generous, careful reader that, by engaging in detail with the sources of Young's work as well as later critics, leaves us with a more theoretically grounded, rigorous and sophisticated account of Young's notion of "structural injustice," and her distinction between "moral" and "political" responsibility. To give just one example, McKeown discusses Hannah Arendt's distinction between moral and political responsibility, as well as contemporary conceptions of virtue ethics, to argue that both what Young calls "political responsibility" and what she calls "moral responsibility," are forms of moral responsibility (McKeown 2024, 165). However, moral responsibility in this sense is other regarding, unlike Arendt's morality, which is self-regarding. Arendt's moral principles are about the self; the moral demand is reflected in the "will I be able to live with myself?" question, and thus sharply distinguished from political considerations, which regard "the world.". This understanding of morality as other-regarding makes it hard for Young to distinguish between moral and political responsibility, which is why she falls on the distinction between backward-looking and forward-looking responsibility to explain the moral/political responsibility distinction, but also why doing so leads to problems (McKeown 2024, 145-155).

Second, *WPCR* advances its own conception of structural injustice and responsibility for it in dialogue with this nuanced version of Young's account and other conceptions of structural injustice. In doing so, it advances our discussion of structural injustice at the theoretical, conceptual and practical levels. One of the most notable features of the book is its ability to ground the theoretical arguments on accounts of actually existing political problems, accounts that reconstruct the legal, political and historical details of an actually existing case rather than relying on imaginary, abstract examples that moral and political philosophers often discuss. My comments will focus on this second line of inquiry, although I think that the contributions to the first are equally important.

At the theoretical level, *WPCR* argues that we need a better conception of structure and structural change than the one that informs Young's account and, under Young's influence, current discussions of structural injustice. The problem with Young's account is that the four elements in her conception of structures – objective constraint, social positions, structures produced in action and unintended consequences (Young 2011, 52-64) – are not well integrated in a

unified theory. In particular, it is unclear what difference, if any, structural positions and structural constraints (constraints that agents experience in virtue of their structural positions) make to Young's view of structural change and, consequently, to her view of responsibility. Young's account of structural change seems to rely exclusively on the notion that structures are produced and reproduced in action, an idea that Young adopts from Anthony Giddens' structuration theory. This exclusive reliance on structuration, McKeown argues, explains why Young assigns the same responsibility to agents that occupy different positions: given that everybody's actions reproduce structures to the same degree, everybody has the same responsibility to change the structures: a shared responsibility, based on their similar contribution to structures, to join collective action (Young 2011, 111-112).

But this cannot be true, McKeown argues. Powerful agents have different responsibilities than "ordinary individuals" because they are in a better position to change the injustice of the structure: they have access to more resources and more "elbow room." to make changes (McKeown 2024, 32-33, 36-37). Young's failure to theorize structural position and integrate it in an account of structural change prevented her from seeing that differently positioned agents have different responsibilities because they have different levels of power. Integrating a theoretical account of structural positions in an account of structural change would have allowed her to analyze "the vested interests that come with various social positions and how this affects agents' behaviour in those positions, and how it generates *power* for particular agents." (28) Moreover, by relying on structuration theory, Young's theory inherits its main problem: it cannot explain change (28). In short, the theoretical shortcomings of Young's account prevent it from offering a good account of structural change, one that does not underestimate the role of powerful agents in bringing about change.

To overcome these theoretical shortcomings, McKeown argues, we should ground our understanding of structural injustice on Margaret Archer's critical realist conception of structure and structural change. On the critical realist view, structures evolve through the interaction of structure and agency (24). Thus, at time T1 there is a structure that conditions (but does not determine) agents' actions. T1 is followed by a period, T2 to T3, of interaction between structure and agency, followed by "structural elaboration," i.e. a new structure, at T4. Structures predate and condition agents' actions, even those that transform them. Structures are characterized by structural positions that endow those who occupy them with different vested interests. Agents in powerful positions have a vested interest in maintaining the structure, while subordinated

agents have an interest in changing it. Each structure comes with a particular distribution of vested interests, knowledge of and attitude towards the structure. The interaction between actions and structures is based on this distribution and results in a new structure: this is the moment of *structural elaboration* (33-34). The process of *structural interaction* is not deterministic; "agents have agency to decide how they will act within structures and contest them" (36).

One of McKeown's main claims is that in this process of structural interaction powerful agents are in a better position than ordinary individuals. They have greater resources and therefore have more 'elbow room' to act both in ways that maintain the structures and in ways that transform them (36). Therefore, they bear a larger share of responsibility than ordinary citizens. They should lead the action for changing unjust structures.

Once we adopt Archer's critical realist conception of structure and structural transformation and consider the role of powerful agents, McKeown argues, we can see that, rather than "the unintended outcome of 'benign social processes'," as Young claims, structural injustice is often deliberately maintained or at least not avoided by powerful agents who would have the ability to eliminate it (41-45).

McKeown distinguishes between three types of structural injustice: deliberate, avoidable and pure. Structural injustice is *pure* when all agents are constrained to such an extent that they cannot avoid but reproduce the injustice, and the injustice is the unintended consequence of their actions (41); this is structural injustice as Young understands it. There may be cases of pure structural injustice (climate change might be a candidate), McKeown argues, but they are not the rule. Structural injustice is *avoidable* when some agents – the powerful ones – are not constrained to such an extent as to not be able to change it; it may be unintended, but it is foreseeable and avoidable (43). Finally, structural injustice is *deliberate* when powerful agents deliberately maintain it (44), as multi-national corporations (MNCs) maintain the powerless and destitute position of workers in sweatshops.

In the case of deliberate and avoidable structural injustice, powerful agents bear moral responsibility to address the injustice because they, unlike "ordinary citizens," have the resources and capacity to effect structural change (36). McKeown makes this case through the example of MNCs in the clothing industry and, in particular, the Bangladesh Accord on Fire and Building Safety signed in the aftermath of the Rana Plaza factory collapse by 190 corporations from 20 countries (80-89). This accord, "a legally binding agreement over five years," that "required signatories to agree to independent inspections to facto-

ries, to remedy any faults and to provide fire and building safety training to staff," and "included Bangladeshi and international trade unions, and was overseen by an independent representative from the ILO" (84), shows that MNCs have the capacity and resources to remedy structural injustice.

In what follows I raise some concerns with two elements of this account: the notion that "powerful agents" should make structural changes and the typology of structural injustice.

There is something immediately intuitive and, I would think, almost uncontroversial in the idea that those who profit most from (and deliberately maintain or can remedy) the vulnerable position of others should be held responsible (morally and politically) for not remedying it. Yet, it is unclear what it amounts to theoretically: What does the example of the Bangladesh Accord show about how change happens and where the power to effect change lies?

As McKeown's detailed account of this example reminds us, in the wake of the 2013 Rana Plaza factory collapse, intense media attention and NGO mobilization "*forced* the global garment industry to face up to the appalling conditions of garment workers" and consequently to sign the Bangladesh Accord (84, my emphasis). While McKeown takes this as evidence that MNCs have the power to effect structural change because they can make the changes that address the injustice, I think it shows the opposite. It shows that change required other agents in the system – NGOs, trade unions, a certain public – to put pressure on MNCs. MNCs' capacity to make this change was the same before and after the Rana Plaza factory collapse. It is, so to speak, a constant, and a constant cannot explain change. What changed – and therefore could explain the change – is the pressure on MNCs from other agents. If anything, the example shows that other agents, not MNCs, exercised their power to bring about change.

If, with Dahl, we think that "*A* has power over *B* to the extent that he can get *B* to do something that *B* would not otherwise do," (Dahl 1957, 202-3, cited in McKeown 2024, 48), then the Bangladesh Accord is a straightforward case of *agents other than the most powerful in the system* getting MNCs to do something they would otherwise not do; on this conception of power, *B*, the agent over whom power is exercised, is *capable* of doing the *X* which they would not do in the absence of *A*'s power. Like *B*, MNCs were *capable* of making the changes they eventually made by signing the Bangladesh Accord, but they would not have made these changes were it not for the pressure put on them by NGOs, trade unions, global publics, etc.

In short, while the example shows that MNCs had the capacity to effect changes, it does not show that the changes came about as a result of their power. On the contrary, it shows that, in spite of their power, MNCs were

forced to bring about change, change that came about as a result of the power of the less powerful agents in the system. This is a case in which change came about in spite of, not because of, the power of the most powerful agents in the system, agents that were forced to do something they would not otherwise have done. It is a case of the powerful agents being under power of the less powerful agents.

This may sound paradoxical if not downright contradictory. How can the powerful agents be under power of the less powerful agents? Does this ignore or reject McKeown's argument that MNCs corporations have systemic power and, in virtue of it, dispositional and episodic power (McKeown 2024, 49-50, 52-67)?

I think the contradiction is only apparent. It disappears if we distinguish between two senses in which we can talk about power: power in the structure and power to change the structure. An agent is powerful in the former sense if it occupies a hierarchically superior, or privileged position in a structure, position that confers benefits or advantages to its occupants. In contrast, agents have power in the latter sense when their actions can effect structural change. McKeown's account conflates these two senses of agents' power. It assumes that those powerful in the former sense – of occupying a position of power in a structure – will have power in the latter sense – will be in a better position to effect change through their actions. But there is no *a priori* reason to think this is always the case. There is no reason to think that those in privileged positions in a structure are necessarily in a better position to change the structure than those occupying the disadvantaged positions. Men, to give just an example, are not necessarily in a better position to change sexist structures than women. Sometimes the structurally privileged may be in the best position to effect change, sometimes they may be in the worst position, and in other cases they may be in the same position as the disadvantaged. Different structures are likely to be different in this respect. To determine in each case which agents are in a position to effect change will require empirical attention to the socio-structural reality of the mechanisms of change for each structure.

McKeown suggests that having power in the former sense (in the structure) puts more resources in one's hands, which gives one power to change the structure. She does not offer an extensive argument for this idea, but the thought seems to be that the powerful will use the resources they have in virtue of their position to maintain their position of power; therefore, they could also use these resources to undermine this position.

There are, however, two problems with this thought. First, resources – or the power one has in virtue of one's control of resources – are not enough to

make one an agent of change. In addition, one has to be the type of agent that, in spite of occupying a position of privilege in the structure, is not (fully) invested in the maintenance of the structure. Secondly, it is not the case that the structurally subordinate lack resources simply in virtue of their subordinate position. Let me take these two claims in turn.

First, one aspect of Archer's critical realist view that recommended it to McKeown is that it allows us to theorize structural position and in particular how an agent's structural position invests them with interests that affect their behavior (28). While undoubtedly MNCs have systemic power, that power accompanies their structural position, in virtue of which they have a vested interest in maintaining, not changing the current functioning of the structure. It is true that powerful actors have "elbow room" in the sense that they are not fully constrained by their position; their position does not determine their actions; there are many actions they can take from within their position. This, however, does not show that those actions open to them are the same as those necessary to effect structural change. Even in those cases in which actions that could result in structural change were open to powerful agents, we are not given any reason to think that they would actually take those actions (without other agents exercising power over them). In short, showing that powerful agents have some "elbow room" in their actions, i.e. some possibilities of action unconstrained by their structural position, does not yet show that their actions will be part of how change comes about. What we need, in addition, is a reason to think that they would take the actions that would modify the structure that benefits them, actions that would undermine their position of privilege.

My argument is not that this reason can never be provided. For some structures it may very well be Rather, my argument is that Archer's critical realist account does not seem the right framework to generate such reasons, in as far as it emphasizes the vested interests of powerful agents to maintain the structure. Moreover, McKeown's discussion of the Bangladesh Accord does not yet provide such a reason. On the contrary, McKeown own analysis provides us with reasons to think the opposite; that is, to think of the Bangladesh agreement as a set of actions that MNCs took to maintain their power-inside-the structure in the face of a challenge to that power. Although under the initial pressure MNCs made changes to some important elements of the system, the agreement had elements that limited those changes: the agreement was limited to five years (it expired in 2021 after a three-year extension), did not include provisions regarding wages or child labor, and it did not upheld rights of collective bargaining (84-87). "In some ways," McKeown argues, "it was a smokescreen promoting an image to ill-informed consumers that brands are

doing more and doing better, when they continued to act so as to perpetuate structural injustice and to use it for their own gain" (87). If this is the case, then we should not interpret this case as one in which there was even incremental, limited change, the type of change that, McKeown argues, should not be underestimated (189). Rather, it would be reasonable to understand these actions as very limited changes for a limited period of time until the media attention moves elsewhere, that is, as ways of maintaining their position of power under conditions of public pressure to change. The second problem with the thought that the privileged have more power to change structures because they have access to resources the subordinate are deprived of is that it assumes that the subordinate lack the relevant resources. This is a mistake because all occupants of structural positions arguably control some resources and, as I argue elsewhere, this control of resources gives the subordinate – understood as a collective, not as individual agents – the power to change the structure in virtue of which they have those resources when they act in ways that do not conform to their mandated use in the structure (Marin 2024).

McKeown agrees that the subordinate position of victims of structural injustice does not deprive them of agency. While they are victims of structural injustice in the sense that they do not benefit from the injustice and they are "powerless in relation to, and rendered vulnerable by, structural injustice," "they are also agents" (170). But if the subordinate have agency, i.e. the ability to act, they must also have power (to act). McKeown is right to distinguish this form of "power with," which she calls "empowerment," from that of "powerful agents," i.e. privilege. However, for reasons I developed above, it is unclear why we should ignore this form of power and focus only on the latter in our account of structural change.

Thus far I took for granted McKeown's typology of structural injustice as pure, avoidable and deliberate. In the rest of these comments, I want to raise some concerns about it. In particular, I will suggest that accepting it makes structural analysis irrelevant. McKeown introduces the distinction between pure, avoidable and deliberate structural injustice as a criticism of and move away from Young's definition of structural injustice as the unintended consequences of a multitude of social processes that, except for their unintended outcomes, does not exhibit any injustice. As such, it is a welcome contribution. However, I think that the typology is premised on blurring the distinction between structural and interactional injustice and thus risks making structural analysis irrelevant.

While Young is often interpreted to say that, by definition, structural injustice takes place when no agent does anything wrong according to "commonly

accepted rules," I think that Young's claim is rather that if we focused on the wrongdoing of agents alone, structural injustice would remain invisible. An account of the wrongdoing of agents alone would not capture structural forms of injustice. Consequently, eliminating this type of wrongdoing, wrongdoing we can recognize with "commonly accepted rules," would not eliminate structural injustice. To recognize structural injustice we need a different, structural perspective (Young 2011, 47-48, 70-71).

In short, Young's claim is not about whether, as a matter of fact, there is or there is not agentic wrongdoing in structural injustice. Rather, it is that agentic wrongdoing is irrelevant to the structural injustice. Cases of injustice that can be traced to specific agents, and that can be remedied through rules for agents do not need a structural analysis. McKeown's notion of deliberate (and arguably avoidable) injustice shows exactly this: that some unjust conditions some groups of people find themselves in can be traced (fully) to the actions (or inaction) of those who benefit from them. Even if this were the case, it does not show that Young's notion of structural injustice needs to be changed. It only shows that some cases – including sweatshops, which Young thought of as examples of structural injustice – are not in fact cases of structural injustice, but old-fashioned cases of agents violating rules of justice that apply to agents, such as "it is wrong to deprive others of their basic rights (and especially wrong to benefit from such a deprivation)." We do not need a structural analysis (or the notion of structural injustice) to theorize these cases to understand the nature of wrongdoing involved in them or the responsibility of the agents involved in them. Like cases of structural injustice, in these cases the wrong may be mediated by complex social processes, but this alone does not make them cases of structural injustice, because their injustice can be traced to specific agents, and can be remedied by enforcing rules for agents. McKeown's disagreement with Young is then not over the definition of structural injustice, but over whether sweatshops count as structural injustice in Young's sense.

These disagreements only show that McKeown's *WPCR* has advanced the debate on structural injustice and the responsibility it creates for us. It is, I think, required reading for anyone interested in these debates.

Works Cited

Dahl, Robert A. 1957. “The Concept of Power.” *Behavioural Science* 2 (3): 201-15.
Marin, Mara. Published online 2024. “Structural Responsibility.” *American Political Science Review*: 1-15.McKeown, Maeve. 2024. *With Power Comes Responsibility. The Politics of Structural Injustice*. London: Bloomsbury Academic.
Young, Iris Marion. 2011. *Responsibility for Justice*. New York: Oxford University Press.

DOI 10.17473/2240-7987-2024-2-4

From Climate Change to Sweatshop Labor: Do "Structural" Injustices Exist, After All?

Rossella De Bernardi*

Abstract

In her new book, Meave McKeown integrates a systematic analysis of power into Iris M. Young's structural injustice paradigm, providing it with the tools to illuminate different agents' relative capacities to reproduce structural injustices (SI). While much needed, this investigation – along with the moral responsibility attributions for SIs (instead of parts of them) it aims to enable – accentuates a tension central to SI theory. On the one hand, McKeown unapologetically questions the aptitude of Young's original notion, "pure" SI in McKeown's words, to capture any real-world injustice altogether. While Climate Change (CC) seems a candidate, other usual suspects – i.e., global poverty and sweatshop labour – would instantiate, respectively, "avoidable" and "deliberate" SIs. On the other hand, these categories sound like oxymorons, considering that these are still to count as "structural" injustices. If it is distinctive of structural causation that it cannot be reduced to the mere aggregation of all contributing agents' conducts, at least some features of the resulting unjust outcomes should be impossible to lay at any specific agent's feet – no matter how extensive and significant their contribution to the overall process. If so, we should only be able to attribute moral responsibility to powerful agents for some – however big – parts of structural injustices rather than the injustices tout court.

Summary: I. Deliberate vs Pure "Structural" Injustices. – II. Structural responsibility to (corporate) agents? – Conclusion. – Works Cited.

In her new book, *With Power Comes Responsibility. The Politics of Structural Injustice*, Meave McKeown (2024) integrates a systematic analysis of power into Iris M. Young's structural injustice paradigm, providing it with the tools to analyze how different structural positions affect agents' relative

* ORCID: 0000-0001-6666-3331.

(in)ability to reproduce – or counter – structural injustices (SI). One of the book's most original resulting contributions is the distinction between three types of SI: (i) pure, (ii) avoidable, and (iii) deliberate. The critical point distinguishing (i) – Young's original understanding of structural injustice – from both (ii) and (iii) is that, in (ii) and (iii), there exist powerful enough (corporate) agents that – in virtue of the structural positions they occupy and the power they consequently enjoy – have "enough elbow room" to remedy the relevant SIs. Still, such agents either (ii) neglect to do so (possibly out of indifference) or even (iii) actively and intentionally maintain them (because they benefit from the unjust states of affairs staying as they are). Such agents are blameworthy for their failures to remedy injustice or their deliberate perpetuation of it. Global poverty is identified as (ii) an "avoidable" SI, whereas sweatshop labor is (iii) a "deliberate SI." Climate change is "tentatively" suggested as a form of (i) "pure" SI (104).

One of the book's central targets is a hasty response that may be tempting when facing injustices arising out of large-scale and multi-agential causal processes under the influence of Young's "social connection model" of responsibility for SIs. Simplifying, faced with the daunting task of identifying culprits in circumstances of SI, we may reach for a one-size-fits-all response, i.e., all agents who contribute to the reproduction of the injustice share in a distinctively *forward-looking-only* form of "political" responsibility to remedy it. Attributions of moral (i.e., backwards-looking) responsibility for having caused the relevant injustices should be sidelined, along with the related blame.

The book's central analysis of the "dimensions and forms" of power provides important analytical tools to make us pause before embracing any such interpretation of the relationship between SI and responsibility (48-67). Compellingly, Mckeown argues that power disparities must be accounted for when *describing* how SIs come about and *normatively evaluating* who is culpable and blameworthy for sustaining them. Specifically, I am highly sympathetic with the idea that such agents as multi-national corporations (MNCs), occupying privileged positions in the structures of the global economy, enjoy power in ways that set them apart from ordinary (i.e., relatively powerless) individuals when it comes to attributing responsibility for SIs.

That said, I will attempt to articulate an underlying tension I perceive in the book. The author's explicit commitment to avoiding reductionism in explaining socio-structural causation seems in tension with the idea that we can correctly identify any (corporate) agents as *the* culprits for the (re)production of structural injustices in a way that can consistently allow us to lay at their feet the responsibility "for the injustice *itself*, not merely some instances of harm" (46,

emphasis added). In a nutshell, here's my doubt: in cases of "deliberate SI," the relevant culpable agents should be sufficiently powerful that the wrong-making features of the injustice can be *causally* linked back to *their agency*. If this is the case, though, is this not *ipso facto* a denial of the hypothesis that the relevant injustice is indeed "structural" – i.e., the result of structural causation – in the first place? In what sense does deliberate SI still count as "structural"? [1]

To shed light on such a tension, I will single out two aspects that an analysis of structural injustice – as in "structurally caused" – can emphasize. At the same time, I will highlight how, when working through the examples, the book seemingly shifts between different characterizations of what makes the relevant cases unjust.

I. Deliberate vs Pure "Structural" Injustices

Let me briefly analyze the ideas of "deliberate SI" and "pure SI" through the real-world cases presented in the book to try and flesh out in some more detail the tension in which I am interested.

Cases of deliberate SI would be such that (44):

> all agents are constrained, but there are powerful agents who have enough room to manoeuvre to change the [unjust] situation [… Still, these] powerful agents want to maintain the vulnerability of the disadvantaged in order to continue to exploit them.

Pure SI, differently, is one in which (41):

> All of the actors are constrained to the point where it is very difficult for them not to participate in reproducing the injustice, and the consequences of their actions are unintended.

Based on such definitions, the three points to make sense of SIs – and distinguishing its types – are:

a. the presence of structural constraints on agency;

b. the actual space left by structural constraints to autonomous agency (meant merely as the opposite of "structurally determined agency");

c. the intention to reproduce injustice through one's agency.

[1] For mere reasons of space, I will focus only on the contrast between "pure" and "deliberate" SI throughout the text, setting aside the contrast with "avoidable" SI.

The two definitions share point *a.*, i.e., the presence of structural constraints on agency (on *b* and *c,* the two notions diverge). Crucially, *a* seems to be the key defining feature of what makes an injustice "structural" in McKeown's account. Or, at least, it seems to me the one given greatest weight throughout the book.

Indeed, in keeping with Young's standard paradigm, one way to make sense of SI is to focus on how it results from causal processes in which most agents, most of the time, *can't help* but be implicated. At least, they can't help it unless they accept to bear heavy costs such as being unable to – or severely hindered in – pursuing the most mundane of their daily tasks. Andreas Malm's (2016, 18-20) characterization of the structural nature of the "fossil economy" offers an excellent picture of such a quasi-coercive nature of structural processes and why McKeown indeed suggests that CC constitutes perhaps the most plausible example of a "pure SI:" [2]

> The fossil economy has the character of a totality, a distinguishable entity: a socio-ecological structure, in which a certain economic process and a certain form of energy are welded together. […] A person born today in Britain or China enters a preexisting fossil economy, which […] confronts the newborn as an objective fact. It possesses real causal powers – most notably the power to alter the climatic conditions on planet Earth, but this only as a function of its power to direct human conduct. A factory manager will be pressured to obtain energy by plugging into the current from the nearest coal-fired power plant rather than building her own waterwheel. The company owner will send her commodities to the world market on cargo vessels, rather than sailing ships. A cashier may have little choice but to commute to the supermarket in a car […].

The relevant point here is that one aspect specifying "structural" in SI is that *all* agency involved in the process causing the unjust outcomes is subject to structural constraints. For many, this fact usually means a lot of pressure and "duress" to act in ways that will (collectively) set conditions of injustice for some. As Malm's vignette shows, none of the agents can easily withdraw their contribution to climate change (hereafter, CC) if they want, minimally, to keep their jobs (that they probably need to pay their bills so that they can turn their heating on and warm up in Winter, an act that, in turn, generates further carbon emissions, and so on).

[2] In fact, McKeown clarifies that one can plausibly hold that from the 80s onward CC has become an avoidable SI since the growing scientific consensus about its reality means mitigation efforts should have started at that point.

Note two important points. First, from such a perspective, feature *c* above – i.e., whether agents fail to act to change injustice or even intend to perpetuate it – can be consistently given a central space while still moving within a properly "structural" analysis of injustice. The approach would be concerned with SI to the extent that all agents are recognized as structurally constrained. Still, it is theoretically consistent – and practically plausible – that we can identify agents who are not structurally constrained in a way that exhausts their space for autonomous agency concerning whether to perpetuate the injustice. If they do perpetuate it, we have a clue to infer their (culpable) intentions (especially if we also know that they benefit from the injustice staying as it is). A related second point is that – and, indeed, this is in line with a suggestion McKeown advances in the book – we can come to doubt that "there are, in fact, any cases of pure structural injustice" (41). The analytical focus from this first perspective is entirely on agency – i.e., on the extent to which different social actors can resist structural pressures and act autonomously.

There is a second dimension to emphasize in characterizing an injustice as "structural." That is, we focus on how essential features of why the outcome turns out to be unjust are altogether irreducible to wrongful agency (Sankaran 2021a, 2021b). What causes injustice is the structure itself. It is to such an aspect, I suspect, that Malm gestures in the passage quoted above when writing that the fossil economy itself – "as a socio-ecological structure ... possesses real causal power." While the power of the structure will always be "a function of its power to direct human conduct," it works and leads to specific (unjust) outcomes *without the need for the contributing agency of any particular actors* (though, of course, there will be more and less significant contributions). Typically, this happens because the number of agentic contributions, so to speak, is very high. That is, the number of contributing agents can be itself a structural property of the causal process. The injustice only originates whenever enough agents contribute. Indeed, size considerations tend to be paradigmatic in explaining how structures have something as an autonomous causal power.

To illustrate, think of Jewish communities' relative vulnerability to religiously motivated violence in different medieval European states (Sankaran 2001a 455-457). Significant are "state capacity" – states' size-related ability to collect taxes and provide services – and weather events leading to small crops and consequent risks of famine. The idea is that we can explain how Jewish communities *in smaller states* were in positions of *greater vulnerability* to violence compared to communities in bigger states – following droughts and small crops – *while still holding constant (groups of) agents' ill-will against*

Jews throughout different medieval European countries. As bigger tax revenues enabled bigger states to provide more services, they enjoyed a more stable legitimation, making them less vulnerable to anti-Jewish pressures coming from the Catholic Church and, therefore, better capable of protecting Jewish communities. The core point is that if we tried to explain the degree of vulnerability to pogroms in terms of agents' intentions and quality of will – e.g., populations' intolerant attitudes or ill will from specific small states' rulers passively witnessing violence within their kingdoms – we would miss distinctively structural causes of the social position of vulnerability to religiously motivated violence that Jewish communities occupied in medieval Europe. Critically, important *wrong-making features* of the injustices – Jewish communities' vulnerability to violence – are simply irreducible to agentic properties – e.g., anti-Jewish attitudes; instead, they are linked distinctively to structural factors – e.g., state capacity.

McKeown too emphasizes such a perspective, for instance, when justifying the adoption of a realist critical social ontology to avoid reductionism in explaining structural causation (23-24). Still, from a similar perspective, considerations concerning agents' intentions are redundant to the causation of unjust structural outcomes since it is a definitional point that no agents have it within their own power to cause the injustice to continue or cease to exist.

Of course, we can consistently look at how agents may ameliorate the injustice or, instead, exacerbate it. As McKeown recognizes (46), Young never meant to deny that agential wrongs abound in cases of SIs (whenever agents may refrain from doing something that contributes to the injustice but does it nonetheless out of bad will). It is open to the standard Youngian story to claim that these agents intentionally and wrongfully harm others. Young's key example of the structural nature of vulnerability to homelessness for working-class single mothers in the US is compatible with the idea that certain landlords may use the power at their disposal (whatever its size) to stir away disliked prospective tenants on account of their being working-class single moms, hence, increasing the vulnerability to homelessness of anyone occupying such gender-class-marital status position. Yet, a critical point in Young's analysis is that intentional harms are neither necessary – nor *per se* sufficient – for injustices to occur. In Young's words, "while illegal or immoral acts certainly contribute to structural outcomes, the people who engage in such acts are not the only perpetrators of the injustice. *There are too many other people also involved*" (Young 2011: 95 – emphasis added).

II. Structural responsibility to (corporate) agents?

As anticipated, what I take to be the most original and interesting claim in McKeown's book is the idea that there are genuine SIs in which, nonetheless, we can hold powerful agents morally responsible for the "injustice itself, not merely some instances of harm" – a point "Young would reject." In different terms, it seems to me the book aims to pin *a sort of structural responsibility on (corporate) agents*. This reading, hopefully not too inaccurate, fleshes out the tension: how can we consistently hold powerful agents accountable for more than some agential "harms" if we want, at the same time, to vindicate the "structural" nature of the injustice itself?

As seen above, one sense of what it is for an injustice to be structural implies that some of its wrong-making features are not causally linkable back to anyone's wrongful agency but result from distinctively structural properties of the social causation process. If this is correct, it is hard to square with the idea that we can pin on any agents – no matter how plausibly powerful and ill-intentioned they are – the responsibility for "the injustice itself." Minimally, the number of further agents involved (many of whom are also very powerful in different domains and senses) will also heavily influence unjust structural outcomes in a way that weakens our ability to pin on any specific agents – no matter how singularly powerful – the responsibility for the structural injustice. Such a conclusion holds, I think, even by looking more closely at the very insightful analysis of power central to the book.

McKeown's target seems something like a structuring power that would inhere to certain socio-structural positions (such as MNCs') when highlighting how they can "set the rules of the game" in their sectors (89). For instance, the analysis shows how (certain) MNCs' decisions in the global garment industry significantly affect the opportunity costs of many others involved in the complex, multi-agential process leading to the significant vulnerability to exploitation – and terrible working conditions – of (typically women) sweatshop laborers at the bottom of the global supply chain (e.g., myriad subcontractors, states whose relatively weak economies rely on cheap garment exports, the workers themselves). Still, multiple and varied factors independently sustaining the structural injustice of sweatshop labor are listed as well – e.g., colonial history, gender norms, market deregulations in the neoliberal era, and "rampant consumer culture in the Global North" (87-88).

I do not aim to deny that MNCs should be blamed for the persistence of several wrongful harms throughout the global supply chain – e.g., by failing to impose the respect of at least safety regulations. As the book discusses, there

are serious wrongs that MNCs refrain from remedying, despite their power to do so, in pursuit of profit maximization (80-89). Still, it is unclear how helpful it is to discuss such aspects of what counts as unjust in the garment industry when making the case that MNCs deliberately perpetuate it as an SI. This is partly due, I suspect, to the fact that the injustice of enduring poor working conditions (especially) at the bottom of the global supply chain seems of a different kind compared to the injustice of the vulnerability to exploitation in the same position.

Namely, the injustice of poor working conditions seems easier to remedy *via unilateral decisions* compared to the injustice of vulnerability to exploitation (and, therefore, it may not count as a properly structural injustice altogether, at least, under the second meaning of "structural" recalled above). Differently, the vulnerability to exploitation of women workers at the bottom of the global supply chain results from several factors that hold independently of any actors' specific decisions (for instance, all those distinctively related to gender structures such as being typically primary caregivers and at once vulnerable to sexual exploitation). Again, while we can – and should – hold MNCs *morally* responsible for deliberately resisting attempts at ameliorating several aspects of the injustice of vulnerability to exploitation – e.g., refraining from imposing unreasonable demands on supply turnaround – it is unclear to me we can consistently see them as morally responsible for the injustice *tout court*. This holds, at least, if the injustice we are talking about is the vulnerability to exploitation of women workers at the end of the global garment industry supply chain.

Conclusion

To conclude, I am sympathetic to many of the book's premises and several of its conclusions. I also believe the analysis of power at its core offers important analytical tools to illuminate SIs and formulate judgments of responsibility for different agents contributing to them. What I am uncertain about is how *With Power Comes Responsibility* proceeds to do so, specifically by opting for a *differentiation of types of SIs*. My qualms ultimately concern the internal consistency of the SI paradigm for how it emerges out of such a move, qualms that are reinforced by the author's suggestions that, in fact, there may not be altogether any pure SIs around. Partly, I find myself left wondering why, then, the account of the book stays within such a paradigm. The worry is also that there is much to the idea that genuine injustices can be properly

structurally caused, rather than always inevitably agentially caused, and that recentering the SI paradigm on agency and individual intentions may ultimately impinge on its ability to vindicate such an intuition, that is also one of its most distinctive features.

Works Cited

Malm, Andreas. 2016. *Fossil capital: The rise of steam power and the roots of global warming*. Brooklyn: Verso books.

McKeown, Meave. 2024. *With Power Comes Responsibility: The Politics of Structural Injustice*. London: Bloomsbury Publishing.

Sankaran, Kirun. 2021a. "Structural Injustice and the Tyranny of Scales." *Journal of Moral Philosophy* 18 (5): 445-472.

Sankaran, Kirun. 2021b. "Structural Injustice as an analytical tool." *Philosophy Compass* 16 (10): e12780.

Young, Iris Marion. 2011. *Responsibility for justice*. Oxford: Oxford University Press.

DOI 10.17473/2240-7987-2024-2-5

Political Responsibility and the Forms of Solidarity On Maeve McKeown's *With Power Comes Responsibility*

David Owen*

Abstract

This commentary argues that McKeown's use of examples draw out philosophical commitments of her account that are not explicitly thematized in that account. Developing this argument in relation to her reflections on solidarity, it is argued, illustrates how her account negotiates and overcomes a potential tension between two different conceptions of solidarity.

Introduction

Maeve McKeown's efforts to build on the work of Iris Young and others (notably Catherine Lu and Alasia Nuti) regarding structural injustice is richly illustrated with empirical examples. These are particularly salient to her discussions of solidarity and acting with political responsibility that draws on her insightful use of Wartenberg's situated concept of power. In this commentary, I want to focus on this part of McKeown's discussion in the context of her reflections on political solidarity and acting on political responsibility. I am less interested here in offering criticisms of McKeown's view than in trying to draw out and make explicit some philosophical commitments which I take not to be theoretically thematized in her argument.

* ORCID: 0000-0001-8016-2947.

I. Victims, Counter-Finalities and Decision-Making Power

There is one way in which McKeown's argument stays closer to the ground than Young's own more ambitious proposals. Let me briefly remind us of Young's position.

In her *Dissent* article 'From Guilt to Solidarity' (2003), Young concludes thus:

> Laws and regulatory institutions are less a basis for political responsibility than a means of discharging it. Where it can be argued that a group shares responsibility for structural processes that produce injustice, but institutions for regulating those processes don't exist, we ought to try to create new institutions.

This discussion of political responsibility aligns with Young's previous work on self-determination and global democracy in which she advocates the following kind of arrangement:

> I propose a global system of regulatory regimes in which locales and regions relate in a federated system. These regimes lay down rules regarding that small but vital set of issues around which peace and justice call for global co-operation. I envisage seven such regulatory regimes …: (1) peace and security, (2) environment, (3) trade and finance, (4) direct investment and capital utilization, (5) communications and transportation, (6) human rights, including labor standards and welfare rights, (7) citizenship and migration. I imagine that each regulatory regime has a distinct functional jurisdiction, with some need for overlapping responsibility and coordination. Each provides a thin set of general rules that specify ways that individuals, organisations and governments are obliged to take account of the interests and circumstances of one another. (2000, 267).

Such a global system is directly concerned to address issues of structural injustice in relation to a global basic structure. It may seem that the establishment of such a global system would render redundant the need for the kind of political responsibility that Young theorises through the social connection model since, presumably, a tolerably just global regulatory regime of the kind that Young envisages as covering labour standards would enable the global apparel system to be brought within the ambit of applicability of the liability model in the same way that responsibility for labour standards with North American and European states are articulated within the terms of the liability model. Is the applicability of the social connection model of responsibility thus to be construed simply in terms of the responsibility to create the institutional conditions of application of the liability model?

In her late discussion of structural injustice within the state, Young characterises the subjection to structural injustice of social groups in terms of *positional difference* within social structures, using examples such as disability, gender, sexuality and race, and one of the reasons that Young (2001) defends the use of group categories in empirical research on inequalities is that group-based comparisons can make visible forms of structural injustice. For our current concerns, however, what is important about Young's discussions of positional difference with respect to social groups within a state emerges from her reflections on the *politics* of positional difference. Young's argument concerning the politics of positional difference is that it cannot be aligned simply with the laws and policies of the state:

> Movements of African Americans, people with disabilities, feminists, gay men and lesbians, indigenous people, as well as many ethnic movements, realize that societal discrimination, processes of segregation and marginalization enacted through social networks and private institutions must be confronted in their own non-state institutional sites. While law can provide a framework for equality, and some remedy for egregious violations of rights and respect, the state and law cannot and should not reach into every capillary of everyday life. A politics of positional difference thus recommends that churches, universities, production and marketing enterprises, clubs and associations all examine their policies, practices and procedures to discover ways that they contribute to unjust structures and recommends changing them when they do. … Numerous social changes brought about by these movements in the last thirty years have involved actions by many people that were voluntary in the sense that the state neither required them nor sanctioned agents who did not perform them. Indeed, state policy as often follows action with civil society directed at undermining structural injustice as leads it. (2007a, 85).

The implication of this passage is clear: although it is the case that bringing social processes – such as the global apparel system – under the authority of a regulatory regime would construct a framework of rights and obligations that allow 'egregious violations of rights and respect' to be handled under the liability model, the social connection model of responsibility retains salience in respect of forms of informal discrimination reproduced through, for example, entrenched societal norms concerning gender. The form of responsibility articulated through the social connection model is central to contexts of structural injustice in which the institutional conditions of applicability of the liability model do not apply and is focused in such contexts on instituting a regime of governance characterised by an impartial public authority. However,

it remains a significant mode of responsibility even in contexts where such an impartial public authority exists but is *refocused* in such contexts as a supplement to the liability model and, hence, as directed to forms of social interaction that cannot easily, while retaining liberal freedoms, be addressed directly through state policies and laws.

There is no reason to think that McKeown disagrees with Young's arguments concerning the need to build regulatory institutions; indeed, that she agrees is at least implicit in her critical remarks on the kind of hybrid global governance via public-private partnerships that have emerged in the international realm (208-212). However, McKeown focus is much more clearly rooted in attention to current practices and contemporary struggles with respect to action in the here and now and can provide guidance for ordinary citizens in how to act. I will come back to this issue shortly, but I want first to take up a concern that arises in relation to Young's work and to suggest that it carries over into McKeown's but that her examples give her the resources to address it.

The issue is one that arises as arises as a direct implication of Young's invocation of Sartre's conception of counter-finalities. She writes:

> The actions and interactions which take place among persons differently situated in social structures using rules and resources do not take place only on the basis of past actions whose collective effects mark the physical conditions of action. They also often have future effects beyond the immediate purposes and intentions of the actors. Structured social action and interaction often have collective results that no one intends and which may even be counter to the best intentions of the actors. Sartre calls such effects counter-finalities. (2007b, 170)

McKeown rejects the understanding of social structures in terms of 'rules and resources' that Young invokes here, preferring the critical realist view advanced by Archer, but this does not, I think, have any implications for the phenomenon that Sartre identifies. The reason that this Sartrean point matters is that it applies equally to any and all efforts to transform a set of social processes in a more just direction.

When we consider the question of how to address structural injustice, Young argues for the epistemic and efficacy benefits of including sweatshop workers in discussions about what to do:

> Victims of injustice have the greatest interest in its elimination, and often have unique insights into its social sources and the probable effects of proposals for change. This point certainly applies in the case of labor conditions in the apparel

> industry. ... Analysts of some strategies in the movement to improve conditions find that they have ineffectual or paternalistic because the workers point of view and active participation have not been properly included. (2006, 185)

Mckeown also stresses this point (230) but she goes further, arguing that 'successful interventions in structural injustice will be grounded in the concerns of the victims and potentially designed and shaped by them too ... The role of the privileged is to support them in an appropriate way' (230-31). Earlier McKeown highlights the example of United Students Against Sweatshops (USAS) noting:

> It is through direct collaboration with sweatshop workers and local unions that they get their information, develop demands and design campaigns. For instance, it is due to worker empowerment that the US anti-sweatshop movement has generally avoided boycotts, since the workers were clear that they wanted to keep their jobs, and boycotts would undermine that. (215)

The importance of this example is that it points to the fact that acknowledging the problem of counter-finalities raises a crucial issue concerning decision-making power *in* the discharging of responsibility that is not simply epistemic or strategic, there is a further and more fundamental reason for those who are most dominated and disadvantaged by a given set of institutional practices and social processes to have a *pivotal* role in determining the courses of action to be taken in transforming these processes, namely, that they are most vulnerable to well-intentioned actions producing unintentional negative outcomes. In other words, we need to reflect on location of persons within structures of decision-making power with respect to social transformations. Thus, on this account, whereas we might argue that *responsibility for change* should be distributed on the basis of the degree of causal role, advantage accrued and power to transform, *decision-making power fo*r *(the direction of) change* should be arranged in terms of degree of subjection to structural injustice. To recall an earlier point from Young's discussion of social groups subject to oppression in *Justice and the Politics of Difference* (1990), this might take the form of allowing sweatshop workers a (qualified) veto with respect to decisions of the anti-sweatshop movement: hence, no boycotts that might put the worker's jobs at risk.

McKeown does not, I think, thematize this issue explicitly but her examples draw attention to it and, as we will see, it matters for how she construes political solidarity. But before we turn to the issue of political solidarity, let me return to the point of contrast between Young's more ambitious project of

proposing a global system of regulatory regimes and McKeown's more local project of offering guidance concerning what to do here and now. The issue here is that systems of structural injustice interact; changes directed at, and successfully addressing some part of, one such system may have negative effects on another. This is salient for issues of political solidarity (as we'll see shortly) but arguably it also indicates the need to identify an overall end – such as a global system of regulatory regimes – at which specific reforms can be strategically directed.

II. Political Responsibility as Political Solidarity

Drawing on Young, Dean, hooks and Medina, McKeown sketches an account of reflective political solidarity in which the work of solidarity entails learning ways of seeing others as equals that are attentive and responsive to difference, to working across difference. But different moments in McKeown's argument seem to point to different conceptions of solidarity across a fundamental distinction, namely, whether these are symmetrical or asymmetrical conceptions of solidarity. Thus, on the one hand, McKeown argues, following bell hooks' criticisms of the feminist movement for its condescending unequal treatment of black feminists that to 'engage in political solidarity means recognizing the oppressed as equal active participants in movements to undermine the injustice that affects them' (p.175) Here the concept of political solidarity is symmetrical in the sense sketched out by Sangiovanni (2015):

I act in solidarity with you when:

1. You and I each (a) share a normatively justified goal (b) to overcome some significant injustice;
2. You and I each individually recognize our responsibilities to do our part in achieving the shared goal in ways that mesh;
3. You and I are each individually committed (a) to the realisation of the shared goal and (b) to not bypassing each other's will in the achievement of the goal;
4. You and I acknowledge our obligation (a) to incur significant costs to realise our goal if necessary; and (b) to share one another's fates in ways relevant to the shared goal.
5. Facts 1.-4. need not be common knowledge.

On the other hand, McKeown also argues, as we noted earlier, 'successful interventions in structural injustice will be grounded in the concerns of the vic-

tims and potentially designed and shaped by them too ... The role of the privileged is to support them in an appropriate way' (230-31). Here the concept of solidarity is construed as asymmetrical. The first view of political solidarity as *shared action* grounded on a *shared goal* is contrasted with a different *shared action* conception of solidarity as '*reason-driven action on other's terms.*' (Kolers 2016: 57, see also Scholz 2015, 273) It is a notable feature of this account that it offers a direct challenge to Sangiovanni's symmetrical view:

> Sangiovanni assumes that solidarity is a *symmetric* relation, such that S is in solidarity with G iff G is in solidarity with S. But solidarity is not a symmetric relation; it is deferential. ... Solidarity is therefore asymmetric ... (Kolers 2016, 61).

For Sangiovanni, solidarity is a symmetric relation because those in a relationship of solidarity have a shared goal and are committed to '(a) to the realisation of the shared goal and (b) to not bypassing each other's will in the achievement of the goal'. By contrast, Kolers argues that solidarity is an asymmetric relation in which S defers to G's specification of the goal and S is committed to not bypassing G's will.

McKeown's first example of the feminist movement seems to align with Sangiovanni's type of view, whereas her second example of the privileged supporting the victims of structural injustice appears to align with the Kolers' type of view. These different examples, however, point to a way of dissolving the apparent conflict between these views. Sangiovanni offers a conceptualisation of solidarity that addresses relations between members of a structural group (e.g., the feminist movement), hence the symmetrical character of his account, whereas Kolers offers a conceptualisation that address relations between members of a group who stand in relations of privilege to the structurally disadvantaged group and the members of the structurally disadvantaged group, hence the asymmetrical character of his account. Rather than seeing one of these as basic to the concept of solidarity in a way that excludes the other, we can sensibly be guided by our ordinary use of the word 'solidarity' to describe both types of relationship and that we see each view as picking out a distinct kind of solidarity: Sangiovanni's account captures the normative character of *solidarity between* members of an oppressed or disadvantaged group: *in-group solidarity*, while Koler's account addresses the normative character of *solidarity with* disadvantaged groups by those who occupy privileged positions as a result of structural injustice: *out-group solidarity*.

This set of distinctions matters for McKeown's argument for two reasons. The first is that, drawing on the work of Ackerley as well as hooks, she high-

lights the importance of intersectionality for theorizing solidarity, but to engage seriously with intersectionality requires engaging both symmetrical and asymmetrical conceptions of solidarity because individuals within a structural group occupy different positions relative to intersections with other structural groups such that they have responsibilities to stand in relations of *solidarity with* (out-group solidarity) and *solidarity between* (in-group solidarity) others within that group. Take the example of women. At its most general, this identifies a structural group 'women' who are disadvantaged relative to another structural group 'men' and in which members of the group 'women' have responsibilities of solidarity towards each other (in-group solidarity) and in which 'men' have obligations of solidarity towards 'women' as a structural group (out-group solidarity) not least as one key dimension of discharging what McKeown, following Nuti (2019), sees as their structural debt towards women. At the same time, however, there are many 'women' who are also members of the structural groups 'Women of Color', 'LGBTQ women', 'Women with disabilities', etc., who have specific *solidarity-between* responsibilities towards each other – in-(sub)group solidarity – and towards whom women who are not members of the relevant structural group have *solidarity-with* responsibilities – out-(sub)group solidarity – that are nested within the wider in-group solidarity responsibilities of women as a general group. This feature of social movements such as feminism that aim to be the self-conscious agency of a structural group, to represent that group as a "group-for-itself", points to the responsibility of women who stand in positions of structural privilege with respect to other women along some dimension of privilege/disadvantage being willing to defer (within limits of justice) to the specific goals set by members of structural group in question. *But* it also points to the responsibility of the relevant sub-group of women being willing to articulate these goals in ways that mesh with the articulation of the wider goals of women as a general structural group and through this with the particular goals of other sub-groups who similarly engage in such articulation of meshing goals. This "dialectics" of solidarity – expressed in ethical concepts such as 'sisterhood' – mediates the relationship of general and particular interests and, to the extent that the relevant responsibilities of solidarity are acknowledged and acted on, disarms the potential for difference to become divisive that may threaten to undo the movement as a whole.

This last point helps to account for the importance of *ethos* in social movements, of the cultivation of an ethical culture of mutual responsiveness that is attuned to diverse structural standings within the group 'women' against the backdrop of a general in-group solidarity relation grounded in the

structural disadvantage of women as a group in society. It is important because 'responsibilities' of solidarity are such that they cannot be fully specified in terms of determinate obligations but require a dispositional relation to the other in and though which their responsibility is worked out in an ongoing and mutual supportive process. This is what underwrites bell hooks criticisms of white feminist's condescending attitudes as failing to cultivate the egalitarian ethos of respectfulness, where this mode of relationship is not required merely as an instrumentally valuable means for achieving the goals of solidarity, rather it is a constitutive part of solidarity, of standing in a relationship of solidarity to others. This is the point that links the concept of solidarity and its expression through notions such as sisterhood or comradeship to the idea of friendship and, hence, to the centrality of ethos.

McKeown's preferred view of solidarity as a "virtue" may be her way of registering this point and it is clear that her examples and the norms concerning how to engage in solidarity in ways that acknowledge intersectionality and the differential position of persons relative to a given structural injustice requires that we acknowledge the duality and dialectic of solidarity in something like the ways sketched out above. What I am offering here is thus not really a criticism of McKeown's account so much as a supplement designed to make theoretically explicit features that are largely implicit in the practical norms of acting-in-solidarity she draws from Ackerley and the examples of solidarity in practice that she provides. It is part of the virtue of solidarity, we may say, that we know the kind of solidarity required of us in any given context of struggle.

Conclusion

This commentary has tried to demonstrate that McKeown's treatment of examples offers us more theoretically than she makes explicit in her argument. I have suggested first that her sweatshop example highlights a reason for privileging the victims of structural injustice with respect to decision-making power in collective action that is not merely epistemic or strategic. I have further proposed that the same example points to the salience of an asymmetrical conception of solidarity that sits alongside the symmetrical conception of solidarity that McKeown invokes in her earlier discussion of political solidarity as a virtue. This, I suggest, can be explicated in terms of the necessity of both symmetrical and asymmetrical views of solidarity for addressing what McKeown takes to be a central requirement of solidarity as a virtue, namely, its ability to be responsive to both intersectionality and positional difference. Perhaps

this reconstructive work on my part does not match with McKeown's intentions? Whether that is so or not, I hope that engaging in such reconstruction and offering it for comment can help to make explicit the theoretical account of political solidarity to which McKeown is committed.

Works Cited

Kolers, Avery. 2016. *A Moral Theory of Solidarity.* Oxford: Oxford University Press.

Nuti, Alasia. 2019. *Injustice and the Reproduction of History: Structural Inequalities, Gender and Redress.* Cambridge: Cambridge University Press.

Sangiovanni, Andrea. 2015. "Solidarity as Joint Action." *Journal of Applied Philosophy* 32 (4): 335-439.

Young, Iris. 1990. *Justice and the Politics of Difference.* Princeton: Princeton University Press.

Young, Iris. 2000. *Inclusion and Democracy.* Oxford: Oxford University Press.

Young, Iris. 2001. "Equality of Whom? Social Groups and Judgments of Injustice." *Journal of Political Philosophy* 9 (1): 1-18.

Young, Iris. 2003. "From Guilt to Solidarity". *Dissent* 50(2): 39-45.

Young, Iris. 2007a. "Structural Injustice and the politics of difference" in *Multiculturalism and Political Theory*, edited by A.S. Laden and David Owen. Cambridge: Cambridge University Press.

Young, Iris. 2007b. *Global Challenges*. Cambridge: Polity Press.

DOI 10.17473/2240-7987-2024-2-6

Response to Critics

Maeve McKeown*

Abstract

This symposium on my book *With Power Comes Responsibility: The Politics of Structural Injustice* raised many interesting and important points. I divide my response into two sections. First, friendly amendments. In this section, I discuss David Owen's points about the role of counter-finalities, symmetrical vs. asymmetrical solidarity, and an ideal theory of justice. I also discuss Vittorio Bufacchi's points about engaging with more theories of structural injustice than just Young's and saying more about intentionality. In the second section, I engage with the critiques of my book. I divide these into two groups. First, Bufacchi and Mara Marin make a similar point about an ostensible binary between the powerful and the powerless in my argument, which I reject by emphasising the different kinds of power that differently positioned agents have within structures. Second, Bufacchi, Marin and De Bernardi all question whether my introduction of intentionality and agency into my account of structural injustice renders the "structural" part redundant. I argue that it doesn't. Drawing on De Bernardi's interpretation of my argument, I restate my case that structural injustice is characterized by structural causation and structural constraint, but that story is incomplete without an analysis of the ways in which powerful agents operate within and manipulate those structures.

Summary: I. Friendly Amendments. – II. The Role of the Powerful and the Powerless. – II. The Role of the Powerful and the Powerless. – Works Cited.

As I write this, Donald Trump has just been re-elected as President of the United States. A convicted felon, a sexual harasser, a corrupt billionaire, an inciter of insurrection, a climate denier and a wannabe authoritarian demagogue. His control of not only the Presidency but also the House of Representatives and the Senate spells grave danger for the US and the rest of the

* ORCID: 0000-0003-3599-2153.

world. For people concerned about structural injustice, and of course, deliberate repression, this is a dark and frightening time.

Structural injustice surrounds us. Growing up, I felt overwhelming guilt about my contribution to it. How could it be that the world was full of poverty, exploitation, climate breakdown and other forms of oppression and domination and I was somehow tied up in it? But I was aware that this might not be the right response. Personalised guilt was possibly inappropriate and potentially even unhelpful. It was reading the work of Iris Marion Young that helped me make sense of structural injustice and my relationship to it. It's not that I and other ordinary individuals are guilty and blameworthy, but that we share a responsibility to act collectively and politically to try to change it. I was sure Young had hit on something important but the more I engaged with her theory, I couldn't help but notice an elephant in the room.

Sure, I and others who are powerless in relation to these structures are not guilty: we have no control over these structures and we are constrained by them in many ways. But not all agents connected to structural injustice can be so absolved. What about the corporations who profit off of structural injustice and who act in ways to ensure it continues? What about rich states who have the capacity to do something about many structural injustices and simply fail to? What about billionaires, like Trump, who continue to amass extraordinary wealth while others live in dire poverty? There was something missing from Young's account. It seemed important to me to keep Young's insights about the nature of structural injustice and ordinary individuals' responsibility for it, but to think more carefully about the role of power. This is when *With Power Comes Responsibility: The Politics of Structural Injustice* was born.

If my book can add anything to thinking through the implications of this current catastrophe, I suggest it is that political responsibility does not stop at the ballot box. Political responsibility is an ongoing responsibility that we all share to engage in collective action against structural injustice whenever and wherever it is possible for us to do so. Political responsibility entails that we develop the capacity for solidarity to engage in resistance across our many and varied differences. This responsibility becomes even weightier in dark times. As I write at the end of the book, "Apathy is not an appropriate response because apathy leads to total political collapse and paves the way for totalitarianism, as Arendt so astutely observed" (McKeown 2024b, 234). We must keep going.

I am immensely grateful to Vittorio Bufacchi, Rossella De Bernardi, Mara Marin and David Owen for their close and insightful readings of my book. It is impossible to know how a book will be received and their commentaries are

reassuring that the book has added something of value to the structural injustice debate. But they are also stimulating and thought-provoking, raising avenues I hadn't thought of and pressing me on some issues that require further elaboration. I will group my response in terms of two main issues: friendly amendments and critiques. The critiques converge on two main issues: the role of the powerful and the powerless, and what is structural about my version of structural injustice.

I. Friendly Amendments

Starting with friendly amendments, David Owen highlights the role of counter-finalities in structural injustice theory – something which I skirted over in the book – and suggests how and why this concept could be further elaborated. Counter-finalities is a concept Young borrows from Sartre and it refers to the ways in which actions within social structures interact in such a way as to cause outcomes not intended by any of the actors. Owen emphasises that this raises something fundamental about political responsibility. While Young and I highlight the epistemic and strategic importance of centering the victims when tackling structural injustice, Owen argues that they ought to have *decision-making power* because "they are most vulnerable to well-intentioned actions producing unintentional negative outcomes." He suggests the example of giving sweatshop workers a qualified veto with respect to the anti-sweatshop movement's decisions. The victims or the subjected are best positioned to see how interventionist actions will potentially affect them and so they should have the opportunity to direct action, even if "*responsibility for change* should be distributed on the basis of degree of causal role, advantage accrued and power to transform." The idea of some sort of veto strikes me as an important idea in theory, but I do wonder about its effects in practice. For example, since the anti-sweatshop movement is a collection of organisations and social movements, can there be veto power over it? Would it be the sweatshop workers themselves or their representatives, e.g., unions, who would exercise this right? Which workers or unions would have this power? Could political struggles over this procedure stymie action? Perhaps a veto is too strong, but I certainly agree that centering the victims in decision-making about collective action should be prioritised. The mechanism for doing this would vary according to context and a veto might be one way to implement that in the right circumstances.

Owen then discusses how decision-making power impacts my conception of solidarity. On the one hand, I seem to propose a "symmetrical" definition of solidarity, where agents are of equal standing and are committed to achieving a shared goal together. On the other hand, because I want to privilege the victims, I am actually committed to an "asymmetrical" definition, where the action is driven on others' terms. Owen helpfully resolves the tension for me. He points out that the symmetrical view explains *solidarity between* members of an oppressed group (in-group solidarity), whereas the asymmetrical view explains *solidarity with* disadvantaged groups by privileged groups (out-group solidarity). Both are important for considering solidaristic action. He gives the example of the group of "women." In order to take intersectionality seriously, women recognise the in-group symmetrical solidarity of working together for a shared goal; in addition, privileged women need to recognise the out-group solidarity of working with women from different kinds of disadvantaged groups. All of this requires cultivating an ethos of solidarity – "an ethical culture of mutual responsiveness that is attuned to diverse structural standings within the group." As Owen suggests, this idea is captured by my understanding of solidarity as a "political virtue," but in providing this analysis of symmetrical and asymmetrical solidarity and the necessity of an ethos of solidarity, he helpfully fleshes out what the virtue consists of.

One further point that Owen brings up, is that Young provided in earlier work an almost ideal theory of global justice in the form of a global regulatory system, or "decentered diverse democratic federalism" as she called it (Young 2010, 32). Owen suggests that there is a "need to identify an overall end" in order to direct solidaristic action. As he points out, my argument "stays closer to the ground" and doesn't offer any such end-goal. To clarify, one reason why I refrain from that is because of the critical realist ontology I employ in the book. I'm committed to the view that the actions of agents within structures interact in unknowable ways, creating outcomes intended by no-one (the counter-finality issue raised above). Such a view makes it difficult to propose an end-goal because the assumption is that even if agents are pushing for such a goal, their efforts might be modified or cancelled out by others' actions. If I return to my opening comments, an authoritarian ruler might suddenly hold unconstrained power in the most powerful country in the world and stamp out many attempts at progress. Even then, we don't know what this behaviour might unleash in the long-term because any such policies will interact with other policies, with international changes, with social movements, with unexpected events like pandemics or major climate disasters etc. The outcomes are not knowable.

Vittorio Bufacchi points out in a similar vein, that while I have a lot to say about the "structural" in structural injustice, that I have less to say about the "injustice." Instead, I defer to Young on this point, agreeing with her that injustice is the domination or oppression of social groups. Bufacchi takes issue with the idea that oppression is the inhibition of self-development, claiming that "I have never met a person who has fulfilled their personal potential," rendering this an unhelpful way of thinking about injustice. However, that's not what Young nor I mean by oppression.

What's missing from Bufacchi's interpretation is the "systemic" or "structural" component. The idea is that some social groups are systematically or structurally inhibited from pursuing self-development (Young 1990, 37). A classic example is women who have to do the "double-shift" – work plus caring for others and domestic labour – who have no time left for themselves. By contrast, men are uplifted by this work because through women's domestic labour, childcare, emotional labour and carrying of the mental load, men are enabled to pursue their careers, as well as live fulfilling family lives. Women's labour is expended without reciprocation, damaging their careers, personal projects, and mental and physical wellbeing in the process. Young identified five faces of oppression – exploitation, marginalization, powerlessness, cultural imperialism and violence – and I supplement this with material deprivation (poverty) and insecurity in relation to climate change. These forces systematically disadvantage particular social groups, which is what constitutes the injustice.

This definition of oppression might mean that more people are experiencing injustice than on some other account of injustice, but I don't think that is a reason to reject it. For example, all waged labourers are exploited under capitalism, which is the vast majority of the world's population. But the beauty of Young's account is to show that not all waged workers are exploited in the same way nor to the same extent. Women and people of colour, or people in the colonies/former colonies, are exploited in specific ways due to their gender, race or nationality. Furthermore, while all waged workers might be exploited, they are not all powerless. Many higher-paid jobs come with autonomy and authority, but other waged workers lack those things in the workplace and spend their lives taking orders from superiors. Also, even if all waged workers are exploited, they don't all face the threat of violence; that is specific to groups marked out for violence such as women, people of colour, indigenous peoples, immigrants, people from minority religious backgrounds, LGBTQI+ people etc. What Young did is to identify social-structural forces beyond the distribution of resources that structurally prevent social groups

from developing their potential, and that is what she called oppression.

While I think this addresses the question of self-development as too ambitious a goal, there is still a definitional question about self-development and self-determination lingering in Young's work, and by extension my own. In fact, I recently examined an excellent PhD thesis that takes up precisely this point: what exactly does and should self-development mean for structural injustice theorists (Bagadirov 2024)? There is a lot more work to be done on this topic and this is an ongoing and important conversation. Having said all of that, the background assumption in my work is that the end goal is some sort of eco-socialist-feminist framework both domestically and globally. Such a framework would address the seven faces of oppression, as well as domination. But the current book isn't about that and it would require a whole other book (several books?) to outline such an ideal theory of justice. This book was about how we conceive of and tackle structural injustice in the here and now. The scale and ambition are smaller. It is also pluralist. I don't expect everyone to subscribe to the ideal theory of justice that I would subscribe to. But I do think that many people want to eradicate various structural injustices in the world. The difference here might be framed in Sen's terms as the transcendental vs. comparative approach to justice (Sen 2009, 15). The transcendental approach identifies an ideal theory of justice and considers how we get from here to there. The comparative approach simply identifies injustice now and tries to improve upon it. My book is in the latter camp. I believe this is worth doing even in the absence of an agreed-upon ideal theory of justice.

Bufacchi filled my heart with joy by saying that I have "written the book that many of us working in the field have been waiting for at least since 1990, the year when Iris Marion Young's *Justice and the Politics of Difference* was published." I certainly didn't imagine raising to such heights! It's an honour to even be considered in that bracket and I'm very grateful for such a compliment. In terms of friendly amendments, Bufacchi raises the point that I focus on Iris Marion Young's structural injustice theory, which is "slightly disappointing." There are other theorists of structural injustice, notably Marx, Johan Galtung and Newton Garver, and engaging with their work would have enhanced the account. I very much agree.

I often think about the relationship between structural injustice and Marxism. Is structural injustice theory reinventing the wheel? Didn't Marx already diagnose these problems for us almost two centuries ago? Or worse, is structural injustice theory a watered-down, liberal-inflected version of Marxian theory? Marx, of course, didn't use the term "injustice" but I still think

there is a lot of potential in exploring this relationship. My preliminary view is that I do think that structural injustice captures something that is under-theorised in Marx and Marxian theory. The idea that the cumulative actions of individuals and groups interacting within and with structures will result in some form of injustice – "pure" structural injustice, as I call it – and that this can be ignored or manipulated by the powerful, "avoidable" or "deliberate" structural injustice – could still be the case in a classless society. I don't buy the idea that once all public resources come under public ownership or the state withers away that there will no longer be any injustices. Our now deeper understanding of injustices based on social-group difference should make us wary of such a monolithic way of thinking about the alleviation of injustice. Structural injustice theory will remain relevant even when a socialist/communist society has been achieved. But certainly, structural injustice theory and Marx's critique of capitalism have much in common and an investigation into these commonalities is more than worthwhile, both when considering Marx's descriptive, and some of his followers' normative, critiques of capitalism.

Structural injustice theory also has a lot to learn from the works of Johan Galtung and Newton Garver, as Bufacchi points out, and I do engage briefly with Galtung's conception of structural violence when discussing global poverty. And, I might add, there is also much to learn from the work of Sally Haslanger, another theorist of structural injustice whose work I didn't systematically engage with in this book (e.g., Haslanger 2016; 2024). These omissions are due to what Bufacchi describes as "the reverence towards Young [which] can reach levels akin to a religious cult." I am a huge fan of Young, of course, which drew me to structural injustice theory in the first place. Young's work is intoxicating because she provided a razor-sharp critique of the distributive paradigm of justice, diagnosing forms of injustice that escaped contemporary political philosophy and that resonated with many people, as well as deftly moving between genres, themes and topics, but always with an eye on injustice generated by social group difference. She is an icon for many for very good reasons. But I too have been surprised and disappointed by the way that Young's structural injustice theory has been taken up and adopted at face value without further interrogation and critique. I hoped in this book to delve deeper into some of, what I consider to be, the main problems with the structural injustice theory we have inherited from Young and to go from there. But I fully recognise that in doing so I have neglected important work in this area and am contributing to a general marginalisation of these other researchers in structural injustice theory, when we as structural injustice theorists ought to be

engaging seriously with this earlier, or in Haslanger's case, contemporary scholarship.

Another friendly amendment offered by Bufacchi is to suggest that my account of structural injustice would benefit from a theory of intentionality. Since the distinction between pure, avoidable and deliberate structural injustice seems to hinge on intentionality, it is a surprising omission. Engaging with philosophy of action would add the missing puzzle piece. For example, drawing on Michael Bratman, he argues that an intentional action may still have foreseen and undesired side effects; in these cases, an agent will have acted intentionally but did not intend the side effects. This might be applicable to deliberate structural injustice, where it's important to consider not only what the powerful agent intended to bring about, but also what they brought about as a consequence of their actions. In this short response, I can't do justice to the implications of this point. It is an angle that I hadn't considered when writing the book but clearly is a vital area of investigation and one I shall work on further.

So far, I believe all the points raised to be friendly amendments, in the sense that they expand upon rather than critique my account of structural injustice. I could say more about an ideal theory of justice, about solidarity, other structural injustice theories, and intentionality. But the remaining points are more critical. They focus on the role of power in my theory and my typology of structural injustice.

II. The Role of the Powerful and the Powerless

The integration of power into structural injustice theory is the main aim of my book and several authors in this symposium commend that. Bufacchi finds this "the strongest aspect" of my book, however, he disagrees with the way I use it. The first concern is that I seem to employ a binary distinction between the powerful and the powerless. Bufacchi ties this to Young's account of powerlessness as a form of oppression and cites her idea that the powerful have authority whereas the powerless are situated to take orders (Young 1990, 56–58). He disagrees with this idea on the grounds that powerlessness is static, whereas disempowerment is dynamic and shows how power is taken away from someone.

I think there are a few things getting mixed up in this comment. When I discuss Thomas Wartenberg (1990) on power in Chapter 2, I use his situated conception of power to demonstrate how some individuals are positioned

within structures to have power over other agents. The example I take from him is the teacher's power to grade the student. This power can have a significant impact on the student's life. But it does not mean the teacher is all-powerful. The teacher is located in a social alignment whereby this power is backed up by an education system and job market that prioritises qualifications. The teacher is constrained by more senior staff who enforce or can amend the grading framework, as well as a pre-existing institutional framework. So, this example shows how power is situated in social context. Therefore, it's not the same as Young's idea of powerlessness as a form of oppression. In this example, we can see that the teacher has authority over the students and can grade them, which is descriptively a type of power (dispositional power). However, the teacher doesn't set the standards for grading, they might have to meet certain targets or grade on a bell curve, or they might face pressure from higher-ups or donors. In that sense, the teacher might be powerless in Young's normative sense of lacking authority over their conditions of work.

The point of Wartenberg's definition is not to make a comment on oppression, but simply to demonstrate how power operates within the context of structures, and on that point, I think his work is invaluable. Wartenberg's situated conception of power explains how social relationships come with relative power and how the power of any agent is backed up by a social alignment. In the book, I give a further example: a landlord has power over a tenant, e.g. to evict them, and this is backed up by the state, courts and police. It also explains how subordinated agents within power structures can counter power within hierarchies, by creating countering alignments or alternative alignments. I use it to make the point that Marin sums up in her commentary: "Young's failure to theorize structural position and integrate it in an account of structural change prevented her from seeing that differently positioned agents have different responsibilities because they have different levels of power."

The second concern that Bufacchi has is that I follow Wartenberg in equating power with domination. This is problematic, he argues, because power is a dispositional concept and it inflects power with a negative normative valence from the outset. However, I invoke Wartenberg in Chapter 2 to discuss how power operates within structures, then in Chapter 3, I interrogate the concept of power in much more detail. In the latter chapter, I draw on a range of power scholars and argue that there are five dimensions and three forms of power. These are as follows:

Dimensions of power:

1. *A* exercises power over *B* by getting *B* to do what they would not otherwise do (Dahl)
2. *A* sets the agenda (Bachrach and Baratz)
3. *A* keeps *B* ignorant of *B*'s true interests (Lukes)
4. Power constitutes the subjectivity of *A* and *B* (Foucault)
5. Agents acting collectively have the capacity to achieve desired goals (Arendt, Allen)

The forms of power are (following Mark Haugaard 2010):

1. Episodic power – an agent's exercise of power
2. Dispositional power – the capacities an agent has, whether or not they use them
3. Systemic power – the system is structured so as to confer dispositional power on certain agents

All of these are descriptive definitions of power. The normative questions arise when I consider whether or not power is being exercised for good or for ill in different cases.

This typology of power, I believe, is capacious enough to respond to Bufacchi's further concerns about power. He suggests that dominant agents' ability to harm the subordinate is a matter of luck, not power. But I disagree. Dominant agents' capacity to harm subordinates without explicitly threatening or coercing them is a feature of their systemic or dispositional power. They are positioned within structures so that subordinates must work around them. Bufacchi gives the example of Supreme Court judges. Each judge has one vote. But if there is a conservative bloc, it is only when they act together that they can be described as powerful. What I would argue is going on here is that each judge has dispositional and systemic power by virtue of their position on the Supreme Court. But this does not mean they can always exercise episodic power to get what they want. Instead, they might have to work in coalition, so using the fifth dimension of power – acting collectively to achieve a desired goal. This, however, does not deny their dispositional and systemic powers. A Supreme Court judge has the capacity to make decisions that you or I simply don't. They are enabled to do this by a complex and intricate judicial system, and they are backed up by the education system, political system and police.

Bufacchi then says that my discussion of power doesn't tell us much about "the ways in which agents use their power." But, again, I disagree. Chapters 3 and 4 go into depth about how corporations in the global garment industry use

their power to deliberately maintain the structural injustice of sweatshop labour. I also discuss how rich states avoid tackling global poverty, and how fossil fuel corporations have delayed and inhibited action on climate change. Bufacchi suggests I make use of Keith Dowding's insight that agents are powerful when they can change the incentive structures of other agents who fail to collectively mobilize. But I believe this is already incorporated in the third dimension of power – Steven Lukes' (2005) insight that *A* keeps *B* ignorant of *B*'s true interests – and I discuss this through the manipulation of consumers by the global garment industry, as well as the industry's use of PR and corporate social responsibility to manipulate others into believing they are acting in ethical and responsible ways, thus stymieing further regulation.

This raises a worry from Mara Marin, who is concerned that I place too much emphasis on the capacity of the powerful to address structural injustice. She cites my example of the Bangladesh Accord that ameliorated fire and building safety in Bangladeshi garment factories over an 8-year period following the Rana Plaza factory collapse in 2013. She writes about my assessment of the power of multi-national corporations (MNCs) in this situation:

> While McKeown takes this as evidence that MNCs have the power to effect structural change because they can make the changes that address the injustice, I think it shows the opposite. It shows that change required other agents in the system – NGOs, trade unions, particular publics – to put pressure on MNCs in the clothing industry. MNCs' capacity to make this change was the same before and after the Rana Plaza factory collapse. It is, so to speak, a constant, and a constant cannot explain change. What changed – and therefore could explain the change – is the pressure other agents put on MNCs. If anything, the example shows that other agents, not MNCs, exercised their power to bring about change.

There are two issues that I want to bring out from this quote, because I believe there isn't actually much disagreement between us here. First, Marin points out that the power of MNCs to have been able to ameliorate fire and building safety issues is "a constant" so cannot explain the change. I would put this in the following terms: MNCs have the dispositional power to make these changes. Agents can have dispositional power, meaning the capacity to do something, whether or not they act on it. So that capacity is (relatively) constant on my view. Second, she points out that it is when other actors in this system put pressure on MNCs that their behaviour changed, e.g., the role of NGOs, unions, the public etc. I agree. I argue that the role of subordinated agents, or agents in a social alignment with powerful actors, is to put pressure on the powerful in order to make progressive changes. Admittedly, I don't

make this point strongly in Chapter 4, where Marin is taking this argument from, because there I want to highlight how the structural injustices in question came into existence and what type of structural injustice I take them to be according to my typology (discussed more below). But in Chapter 8, I argue that the role of subordinate and socially-aligned agents is to act through civil society to pressure powerful agents for change. I don't believe that MNCs and powerful states can be trusted to make these changes of their own accord, instead, the role of everyone else is to pressure them for change. I argue that we can't just leave it up to powerful agents to address structural injustice because they cannot be trusted to do it, nor to do it right. Instead, other actors, including ordinary individuals, need to act collectively to create alternative alignments or countering alignments. I offer some examples from recent anti-fossil fuel activism. And that is what happened in the Rana Plaza case; subordinate and solidaristic agents, as well as other agents in the social alignment (e.g., national governments) put pressure on the industry to change. So, when Marin says that I privilege power in virtue of structural position over the empowerment of the victims, I disagree. I believe that both work dynamically with each other. It is pressure from below and through the social alignment of the powerful that will push them to change.

Marin continues that the Bangladesh Accord example

> shows that, in spite of their power, MNCs were forced to bring about changes, that these changes came about as a result of the power of the less powerful agents in the system. This is a case in which change came about in spite of, not because of, the power of the most powerful agents in the system, agents that were forced to do something they would not otherwise have done.

This is where I disagree. What I believe the Bangladesh Accord shows is that MNCs had the dispositional power to make the relevant changes, but they only acted on this when they were forced to do so by the pressure of other agents acting in solidarity through civil society. If MNCs didn't have the dispositional power, the activities of relatively powerless agents in civil society would not have been effective. It was essential that MNCs had the dispositional power to make changes for the changes to be made. But Marin disagrees suggesting that, "There is no reason to think that those in privileged positions in a structure are necessarily in a better position to change the structure than those occupying the disadvantaged positions." She gives the example of men in sexist structures; they might not be in a better position than women to effect change. However, this example is not the same as the Bangladesh Accord ex-

ample. Men are not a corporate agent with dispositional power. Men are a diffuse and diverse aggregate. This has two implications. First, they do not have the dispositional power to effect changes in sexist social structures in the way that corporate agents within the garment industry have the dispositional power to make changes in that industry. Second, men, as an aggregate, cannot bear moral responsibility as a corporate agent to make changes. In the Rana Plaza case, subordinate agents were able to pressure powerful agents to make changes that were only within the power of the powerful agents to make, and that they were morally obligated to make. The subordinated agents themselves could not have created a legally-binding Accord. These are two different kinds of cases. Moreover, I emphasise that it is always essential to listen to the victims of structural injustice when making changes. The Bangladesh Accord included trade unions from the outset. Tackling sexism in all its various manifestations would, of course, involve centering the voices of women.

Marin goes on to question whether the Bangladesh Accord was an example of structural change at all. As I point out in the book, it was significantly limited in its remit, not applying to wages or collective bargaining, only applying to Bangladesh, and has now expired. Was it not, then, "a limited change for a limited period… that enabled the powerful agent to ultimately maintain their position of power"? I think it is more complicated than that. The Accord provided a window into what is possible in the global garment industry. *If* there was a global legally-binding agreement that ensured fire and building safety, and if this were supplemented by further measures like addressing a minimum wage and rights to join unions, then we could see structural change in the industry. There would be a human rights floor which is legally enforceable. The Bangladesh Accord provided a glimpse of this possibility and the International Accord is continuing this work (Accord 2022). The UN is also continuing this work by considering whether its Guiding Principles on Business and Human Rights will include legally enforceable mechanisms (Trebilcock 2020). What it shows is that the public pressure required to make this bigger structural change needs to be ongoing and relentless. But of course, as Marin points out, the media spotlight has moved on and the pressure is not there in the same way as it was in 2013. This does not mean, however, that it could not return and the Accord shows us what is possible and what could be aimed towards in this particular industry.

III. Is Structural Injustice Structural or Agential?

Now I want to address the main criticism that comes up in Marin, Bufacchi and De Bernardi's commentaries: what is actually structural in my account of structural injustice? If intentional agential action is a central component of my account, doesn't the whole concept of structural injustice become redundant?

To recap, based on the critical realist social ontology and understanding of power discussed above, I suggest there are, in fact, three types of structural injustice. Pure structural injustice is the kind that Young identified: it's the cumulative outcome of social-structural processes that constrains all agents and requires systemic change. Avoidable structural injustice is also the outcome of cumulative social-structural processes, but those outcomes are foreseeable and there are powerful agents who have the capacity to change it. Deliberate structural injustice is also the outcome of cumulative social-structural processes, but it is deliberately perpetuated by powerful agents because they are benefitting from it.

The authors are concerned that by inserting intentionality back into the concept of structural injustice, that I lose what is distinctive about structural injustice in the first place, and risk rendering it redundant. They press this point in a number of ways:

> "In a nutshell, here's my doubt: if, in cases of "deliberate SI," the power the relevant agents exercise is sufficient to allow *causally* to link the structurally wrong-making features of the injustice back to *their agency* (and their intentions), is not that *ipso facto* a denial of the structurally caused nature of the relevant injustice? If not, in what sense does deliberate SI still count as "structural"?" (De Bernardi 2024)

> McKeown's idea of deliberate structural injustice wants to reintroduce intentionality within the scope of structural injustice. While there are merits to this, there is also the risk that the most distinctive element of this injustice is stripped away. (Bufacchi 2024)

> I think that the typology – and its contribution to the literature – is premised on blurring the distinction between structural and interactional injustice and thus risks making structural analysis irrelevant. (Marin 2024)

I understand the force of this critique and I'm glad they have raised it because it gives me the opportunity to clarify. De Bernardi tries to breakdown the distinction as follows:

"the three points to make sense of SIs – and distinguishing its types – are:

a. the presence of structural constraints on agency;

b. the actual space left by structural constraints to autonomous agency (meant merely as the opposite of "structurally determined agency");

c. the intention to reproduce injustice through one's agency."

De Bernardi argues that what all three versions of structural injustice I identify share in common is *a* – "the presence of structural constraints on agency." She uses a helpful example from Andreas Malm to make this point, who discusses how a person is born into the fossil-fuel economy and can't help but operate within it to live and to work. However, she then makes a second point about what structural injustices share in common, which is not in the above breakdown – "What causes the injustice is the structure itself." She draws on Kirun Sankaran's example of Jewish communities' relative vulnerability to violence in medieval European states. This cannot be explained by agential factors alone, e.g., anti-Jewish attitudes, but instead this vulnerability is bound up with the structure, in this case state capacity. But "If this is correct, it is hard to square with the idea that we can pin on any agents – no matter how plausibly powerful and ill-intentioned they are – the responsibility for 'the injustice itself.'"

De Bernardi applies this to the central case of the book – sweatshop labour. She separates out two issues: the poor working conditions themselves and the position of being vulnerable to exploitation in such a job. The former can be addressed "via unilateral decisions" of MNCs but the latter cannot. So perhaps, she suggests, the former is not really a case of structural injustice at all. The latter is a case of structural injustice but it falls under Young's original framework whereby it is not possible to pin blame on any agents for it.

This is a perceptive point, but I believe my account can respond to it. The poor working conditions are a structural injustice on my view. If we take De Bernardi's two criteria for a structural injustice – structural constraint and structural causation – then both apply in this case. First, everyone is constrained in this system. MNCs are competing in a competitive industry and have to keep costs low. Sweatshop owners are competing with many other factories and have to keep their costs low and turnaround times quick. However, the largest and most powerful MNCs work together through the mechanisms I mention in the book – lobbying, setting industry standards, corporate social responsibility, and manipulation – to maintain the status quo. They could do the opposite – they could work together to improve conditions in the industry, like they briefly did during the Bangladesh Accord under intense in-

ternational pressure (discussed above). The industry is characterized by subcontracting at various and multiple levels. MNCs are not in control of what happens further down the supply chain, but because of the system they have been part of developing, the downward pressure is immense, causing subcontractors to pay workers dire wages under terrible conditions. The pressure is systemic, but the systemic pressure has at least partly been shaped, or at least manipulated, by the activities of powerful MNCs and their home states at the level of global governance.

Second, the possibility for these working conditions to arise in the first place emerged from the complex interplay of decolonization, changes in the processes of production, gender stereotyping and other factors. But this does not mean that there was no manipulation of the structures. In the volume *Structural Injustice and the Law*, I delve into this more deeply (McKeown 2024a). There I show how the garment industry has been highly regulated, starting with John F. Kennedy's international trade agreement on cotton in 1961 and developed into a quota system under the Multi-Fibre Agreement (MFA) in 1974. The aim was to protect the US garment industry from a sudden influx of cheap foreign imports. The global garment industry, as Jennifer Bair points out, was the most regulated industry in the world: since 1961, "apparel production has been among the most protected manufacturing activities in the global economy" (Bair 2008, 3). The EU and Japan also had a quota system under the MFA. These powerful state actors manipulated the industry until this was no longer an option under the WTO. Even then, they eeked out the quota system as long as they possibly could. To me, this shows that both things can be true simultaneously: there can be an injustice which emerges from structural processes but there can also be powerful actors pulling the strings and manipulating the situation for maximum benefit. This is not an analytically neat point. Instead, it reflects the interplay of structure and agency, rather than privileging one over the other, as philosophy often invites us to do. Both are present. What characterizes deliberate structural injustice is that agency is used to manipulate the structures so that they continue to benefit the powerful. I don't believe that this does render the idea of structural injustice redundant, instead I think it more fully reflects the reality of how structures work in practice.

It is a difficult theoretical point to make, which is one reason why I engage in empirical analysis showing how structure and powerful agency interact in practice. In chapter 4, I apply the typology of structural injustice to case studies. In each case, I discuss the genealogy of the injustice, demonstrating that each injustice is the unintended cumulative outcome of social-structural pro-

cesses, but that powerful agents respond to this situation in different ways. In the case of sweatshop labour, powerful corporations and their home states manipulate the situation to their advantage, using international law and bargaining power to ensure that the status quo is maintained and they can continue to profit from oppressive working conditions. In the case of global poverty, even though this has not been caused by any one agent, I argue that there are measures available to powerful states to alleviate this injustice and they simply fail to do it. In the case of climate change, I argue that it is also the unintended cumulative outcome of social-structural processes, but that it is so "baked-in" to the system, in Nancy Fraser's words, that it is a pure structural injustice and requires system overhaul. This is not to say that no agents can be blamed for their contributions to climate change. Fossil fuel corporations can be blamed for their disinformation campaigns, greenwashing, and ongoing attempts to maintain our reliance on fossil fuels. Powerful states can be blamed for dragging their feet on regulation, adaption and mitigation. But ultimately, if climate change is going to be overcome, we have to change our socio-economic system, making it, I suggest, a pure case.

De Bernardi's second point, that being vulnerable to exploitation in a sweatshop job in the first place cannot be the fault of MNCs, is an important one. But I don't think this latter injustice is their fault. Instead, this is the outcome of a different structural injustice, namely poverty. Poverty is gendered and racialized and leaves certain social groups vulnerable to highly exploitative jobs. Poverty, I argue, is an avoidable structural injustice, meaning that if rich states wanted to eradicate (extreme) poverty, they could. There are many solutions on the table, what is lacking is political will.

Marin argues that my analysis of sweatshops shows that it is not a structural injustice at all, "but old-fashioned cases of agents violating rules of justice" so "We do not need the notion of structural injustice to understand the nature of injustice in these cases." I strongly disagree with this point and I hope to have demonstrated above why that is the case. To reiterate, structural injustice is characterized by structural causation and structural constraint. These are fundamental components of the sweatshop case. But the story is incomplete without an analysis of the ways in which powerful agents operate within and manipulate those structures. That is the part that is missing from Young's analysis.

In closing, I would like to express my gratitude again for each of the authors for their careful and enlightening engagement with my work: David Owen, Vittorio Bufacchi, Mara Marin and Rosella De Bernardi. What I think we all agree on is the value of structural injustice theory, the importance of inte-

grating power within it, and the solidarity of the victims and others working collectively to change it. Those are the core issues that I hoped to get across in the book and that remain so urgent in our constantly changing political landscape. Even if we disagree on some of the specifics, I hope that we will all continue to work together to challenge structural injustice and to pressure the powerful for change.

Works Cited

Accord, International. 2022. "About Us". https://internationalaccord.org/about-us.

Bagadirov, Aziz. 2024. "Human Flourishing and Structural Injustice". European University Institute.

Bair, Jennifer. 2008. "Surveying the Post-MFA Landscape: What Prospects for the Global South Post-Quota?" *Competition & Change* 12 (1): 3-10.

Bufacchi, Vittorio. 2024. "Where Is the Injustice in Structural Injustice?" *Philosophy and Public Issues*.

De Bernardi, Rosella. 2024. "From Climate Change to Sweatshop Labor: Do "Structural Injustices Exist, After All?" *Philosophy & Public Issues*.

Haslanger, Sally. 2016. "What Is a (Social) Structural Explanation?" *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 173 (1): 113-30.

Haslanger, Sally. 2024. "Agency under Structural Constraints in Social Systems". In *What Is Structural Injustice?* edited by Jude Brown and Maeve McKeown, 48–64. Oxford: Oxford University Press.

Haugaard, Mark. 2010. "Power: A "family Resemblance" Concept". *European Journal of Cultural Studies* 13 (4): 419-38.

Lukes, Steven. 2005. *Power: A Radical View*. London: Palgrave MacMillan.

Marin, Mara. 2024. "Commentary on Maeve McKeown's With Power Comes Responsibility: The Politics of Structural Injustice". *Philosophy and Public Issues*.

McKeown, Maeve. 2024a. "The Law's Contribution to Deliberate Structural Injustice: The Case of the Global Garment Industry". In *Structural Injustice and the Law* edited by Virginia Mantouvalou and Jonathan Wolff, 82-104. London: UCL Press.

McKeown, Maeve. 2024b. *With Power Comes Responsibility: The Politics of Structural Injustice*. London: Bloomsbury Academic.

Sen, Amartya. 2009. *The Idea of Justice*. London: Allen Lane.

Trebilcock, Anne. 2020. "The Rana Plaza Disaster Seven Years on: Transnational Experiments and Perhaps a New Treaty?" *International Labour Review* 159 (4): 545-68.

Wartenberg, Thomas. 1990. *The Forms of Power: From Domination to Transformation*. Philadelphia: Temple University Press.

Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University Press.
Young, Iris Marion. 2010. "Hybrid Democracy: Iroquois Federalism and the Post-colonial Project". In *Global Challenges: War, Self-Determination and Responsibility for Justice* edited by I. M. Young, 15-39. Cambridge: Polity Press.

Special Sections

The Democratic Containment of Fake News and Bad Beliefs

DOI 10.17473/2240-7987-2024-2-7

The Democratic Containment of Fake News and Bad Beliefs

Enrico Biale* and Gianfranco Pellegrino**

Abstract

The propagation of fake news has given rise to a pervasive sense of apprehension regarding its ramifications for democratic societies. The disruptive influence of digital technologies has intensified the repercussions of information manipulation and eroded the epistemic foundations of democratic deliberation to an unprecedented degree. Consequently, the dissemination of fake news poses a substantial threat to fundamental democratic values such as freedom, autonomy and equality, giving rise to pressing questions regarding the optimal balance between safeguarding these principles and preserving freedom of expression. These issues were addressed in a funded research project "Deceit and Self-Deception: How We Should Address Fake News and Other Cognitive Failures of the Democratic Public" (PRIN 2017), conducted by a team of scholars from four Italian universities (UPO, UNIGE, UNIPV, LUISS). The primary findings of this research were presented in a workshop, which was held at Luiss on 26-27 October 2023. A selection of the papers presented at the workshop have been selected fort his special section of Philosophy and Public Issues.

The spread of fake news, a term that encompasses various forms of misinformation, deceptive narratives and biased reporting, has led to widespread concern about its impact on democratic societies. The relevance of truth and falsehood in a political context has been the subject of much debate among political theorists since the time of Plato, whose works explored the concept of objective moral truths as the basis for good government and the paradoxical nature of permissible lies.

However, Brexit, the 2016 US presidential election, and the spread of right-wing parties have demonstrated the unprecedented ability of fake news to distort public opinion, influence electoral outcomes and exacerbate social polarisation. The disruptive impact of digital technologies has amplified the

* ORCID: 0000-0001-7899-0989.

** ORCID: 0000-0002-8029-3936.

consequences of information manipulation and eroded the epistemic foundations of democratic deliberation to an unprecedented degree.

Scholarly investigations into the epistemology of fake news delve into how fake news interacts with human cognitive biases and the structural dynamics of information dissemination (Bernecker, Floweree, and Grundmann 2021). Fake news often exploits 'cold' cognitive biases, such as the confirmation bias and the availability heuristic, as well as 'hot' motivational factors related to group identity and affective polarisation. These cognitive failures reflect broader issues of epistemic injustice and vulnerability, particularly in digital environments that prioritise virality over veracity and reward tribal loyalty over evidence-based reasoning (Levy 2021).

The consequences of this phenomenon are profound: rather than simply misinforming the public, fake news actively reshapes epistemic norms, thereby undermining public trust in knowledge institutions (Bernecker, Floweree, and Grundmann 2021) and promoting the uncritical acceptance of distorted narratives, thereby undermining collective deliberation (Sunstein 2017; Sperber and Mercier 2017). As a result, the spread of fake news poses a significant threat to fundamental democratic values such as freedom, autonomy and equality, raising pressing questions about the optimal balance between safeguarding these principles and preserving freedom of expression (Galeotti 2018).

These issues were addressed in a funded research project Deceit and Self-Deception. How We Should Address Fake News and Other Cognitive Failures of the Democratic Public (PRIN 2017) conducted by a team of scholars from four Italian universities (A. E. Galeotti, M. Benzi, E. Biale, J. Marchetti, C. Meini, M. Novarese, L. Santi Amantini Università del Piemonte Orientale; G. Pellegrino, G. Floris, G. Sillari, Luiss; I. Carter, F. Liveriero, Riccardo Sportono Università di Pavia; V. Ottonelli, C. Burelli, C. Klijnman, S. Langella, M.S. Vaccarezza, F. Zuolo Università di Genova). The main objective of the project was to examine the numerous cognitive failures of democratic publics and to assess the institutional responses developed to address them, arguing for solutions that enhance epistemic resilience rather than punitive measures that risk authoritarian overreach. The containment of politically dangerous fake news, this perspective argues, must be grounded in and promote liberal democratic values rather than undermine them. Too often, the debate on the disruptive effects of fake news has led to proposed responses in the form of greater control over social media, political information and communication, suggesting epistocratic measures to ensure the proper exercise of democratic participation.

To avoid these shortcomings and to embody liberal democratic ideals, the project developed a different approach, according to which fake news is neu-

tralised by expanding (rather than restricting) the space for freedom of speech and information in a democratic polity.

In order to achieve these goals, the project developed: a) a critical analysis of the cognitive traps that threaten proper democratic life; b) a normative assessment of the proposals that have been put forward to address these problems, questioning both the possible efficacy of these proposals and their potential risks in terms of safeguarding the autonomy of democratic citizens; c) possible remedies to these epistemic failures, including through the development of appropriate cognitive virtues.

a) The first objective involved a careful analysis of the generic category of 'fake news', in order to provide a typology of the various cognitive biases that lead citizens to hold epistemically unjustified beliefs and spread them on the web. The focus of public discourse and academic reflection on 'fake news' has predominantly been on media (both traditional and social) and collective communication processes. In contrast, this project develops a novel perspective that explores the cognitive failures of the democratic public. A central question concerns the intentionality of distortions, whether they are deliberate or not, and the underlying motivations that drive them. Addressing these questions is essential in order to identify individual and collective responsibilities in the production, dissemination and perpetuation of misinformation and distorted beliefs.

b) The project called for a re-evaluation of recently proposed institutional responses to these cognitive failures. These responses were examined in terms of their consistency with the core values of liberal democracy. The analysis revealed that, despite the seriousness of the problems that cognitive traps pose for democratic stability and effectiveness, the proposed solutions may actually be counterproductive. For example, while cognitive traps may compromise individual autonomy of judgement, countermeasures that impose restrictions on access to websites and guided use of the web take the form of vertical acts of paternalism that risk damaging both individual freedom and citizens' autonomy of judgement.

c) The third objective involved a careful consideration of the implications of such cognitive failures for democratic ethics. Widespread alarmism about the collective mechanisms of democratic information distortion is linked to implicit assumptions about individual cognitive requirements for accurate information and the proper functioning of a democratic system. This research challenged these assumptions and advanced a critically informed view of the agential competences and cognitive virtues necessary for the proper functioning of a democratic system.

The three theoretical aims of this research project were complemented by a more applied investigation into the philosophical analysis of climate denialism, taken as a paradigmatic case of democracy-threatening disinformation. The analysis produced a list of factors that underlie climate denialism (complexity of the information base, demandingness of mitigation and adaptation policies, dispersion of agency and causal responsibility in climate change scenarios), a discussion of responsible vs. permissible poor cognitive performance, and a list of non-epistocratic countermeasures to climate denialism.

The main results of this research were presented in a workshop "The Democratic Containment of Fake-News and Bad Beliefs", held in Luiss on 26-27 October 2023, and we decided to collect some of them in this special section of Philosophy and Public Issues. In his paper, Neil Levy challenges the traditional understanding of fake news as inherently deceptive. He introduces the concept of "non-deceptive fake news," which is consumed and shared not for belief but as a rhetorical tool for political signalling and group identity reinforcement. According to Levy, many individuals share fake news not because they believe it but to express loyalty to their political or social groups. This behavior highlights the performative nature of fake news in polarized contexts.

However, the non-deceptive nature of fake news does not weaken its dangerous effects. Even when not believed, fake news disrupts political debates, pollutes the informational environment, and contributes to affective polarization. As a consequence, Levy suggests, strategies to combat fake news should focus on its rhetorical and performative aspects rather than solely addressing belief correction. This perspective broadens the understanding of fake news by highlighting its role in social and political discourse, beyond its epistemic effects. The idea that fake news go beyond epistemology is one of the main premises and outcomes of the project leading to the October workshop.

In her paper, Cathrine Holst examines the relationship between democratic governance and expert knowledge in the context of misinformation. She lists three ideal-typical approaches to democratizing expertise:

1. *Science in Democracy*: This approach emphasizes the independence of scientific institutions and their role in providing neutral, evidence-based input to democratic processes.
2. *Direct Democratization*: Advocates for greater public involvement in expert decision-making, emphasizing transparency, participation, and accountability.
3. *Partisan Expertise*: Recognizes the political and value-laden nature of expertise, challenging the notion of neutrality and advocating for democratic oversight of expert bodies.

Holst argues that misinformation challenges the balance between expertise and democracy by undermining trust in expert institutions. As a consequence, she advocates for "epistemically justified expertise," which integrates epistemic standards with democratic principles to address the dual challenges of misinformation and technocratic elitism.

Säde Hormio delivered a paper focussing on the collective dimensions of misinformation, arguing that bad beliefs are often rooted in "bad epistemic neighborhoods" created by collective agents. Hormio claims that misinformation is often propagated by collective agents such as corporations, governments, or media organizations. These entities bear moral and epistemic responsibility for creating environments conducive to bad beliefs. Moreover, Hormio distinguishes between misbelievers (motivated by ideology) and disinformers (motivated by instrumental goals), noting that the latter are more culpable in a moral sense. Finally, Hormio calls for systemic interventions to improve epistemic environments, emphasizing the role of collective responsibility in addressing the root causes of misinformation.

In her paper, Laura Santi Amantini emphasizes the role of fake news in exacerbating existing epistemic and social inequalities. She identifies two primary ways in which fake news disproportionately harms marginalized groups. First, fake news amplifies epistemic inequality. Fake news disproportionately affects citizens with lower epistemic skills or access to reliable information. These groups are more vulnerable to believing misinformation, which further entrenches their epistemic disadvantages and reduces their capacity as knowers. Second, fake news reinforces social oppression. Anti-minority fake news targets marginalized groups such as immigrants, Muslims, and Roma, perpetuating negative stereotypes and systemic oppression. This type of misinformation denies these groups equal status in democratic societies, undermining political equality and reinforcing cultural imperialism.

Margherita Benzi, Irene Maria Buso, Paolo Chirico, Jacopo Marchetti and Giacomo Sillari presented two papers illustrating the results of the experimental studies conducted in the project. In "Uncertainty and Fake News", fake news is conceptualized as "low-informative signals", showing how it exploits uncertainty to shape belief formation and decision-making. Experimental findings reveal that fake news often aligns with self-interest, leading individuals to favor options that benefit themselves, even when the information is partial or ambiguous. In addition, the ambiguity inherent in fake news allows individuals to justify self-serving decisions while maintaining plausible deniability. This mirrors real-world tactics used in misinformation campaigns, such as

climate change denial or vaccine misinformation. In "Believing vs. Sharing Fake News", two primary explanatory frameworks are identified:

1. **Bounded Rationality**: This approach attributes the acceptance of fake news to cognitive biases and limitations, such as reliance on heuristic thinking (e.g., System 1) over analytical reasoning (e.g., System 2).
2. **Expressive Rationality**: Rooted in identity politics, this framework suggests that individuals share fake news to affirm their group identity rather than out of genuine belief. Sharing serves as a form of social signaling, aligning with expressive loyalty rather than accuracy.

These findings suggest that sharing behavior is influenced by factors beyond belief, such as the desire to mitigate polarization or promote inclusiveness.

Even if they are not a fully comprehensive representation of the outcomes of the project, these articles collectively highlight the main results of it. They emphasize the need to address the epistemic, social, and collective dimensions of misinformation through systemic interventions, epistemically justified expertise, and policies that balance democratic values with epistemic integrity, avoiding any undue restriction of democratic freedoms and rights.

Works Cited

Bernecker, Sven, Amy K. Flowerree, and Thomas Grundmann. 2021. *The Epistemology of Fake News*. New York: Oxford University Press.

Galeotti, Anna Elisabetta. 2018. *Political Self-Deception*. Cambridge: Cambridge University Press.

Levy, Neil. 2021. *Bad Beliefs: Why They Happen to Good People*. New York: Oxford University Press.

Sperber, Dan, and Hugo Mercier. 2017. *The Enigma of Reason: A New Theory of Human Understanding*. Allen Lane.

Sunstein, Cass R. 2017. *#republic: Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press.

DOI 10.17473/2240-7987-2024-2-8

Fake News as Rhetorical Weapon*

Neil Levy**

Abstract

Most philosophers hold that fake news is intended to deceive, either about its subject matter or about the intentions of those who produce it. I argue that a significant proportion of 'successful' fake news – successful in providing political partisans with rhetorical weapons – is not intended to deceive consumers. It is nevertheless relied upon by partisans for political signaling and even to win arguments. Claims need only have a surface veneer of plausibility to be relied on for these purposes: so long as people can present themselves as believing them, they can rely on them in political contexts. When fake news serves this kind of function, it is neither intended to deceive anyone nor does it succeed in deceiving those who rely on it. I argue that fake news can be harmful even when it is neither deceptive nor deceiving: it blocks political debate, it pollutes the epistemic landscape and it leads to affective polarization. If fake news is less often believed than is commonly thought, some ways of responding to it are likely to be less effective than their proponents suggest.

Summary: Introduction. – I. Fake News and Deception. – II. Non-Doxastic Explanations of Fake News. – III. Fake News as a Rhetorical Weapon. – IV. The Harms of Fake News. - Conclusion. – Works Cited.

Introduction

Since 2016 – the year of the Brexit referendum and the year Donald Trump was elected president of the United States – fake news has attracted a lot of public and academic attention. Philosophical debate has centered on its nature

* I am grateful to the John Templeton Foundation (grant #62631) and the Arts and Humanities Research Council (AH/W005077/1) for support.

** ORCID: 0000-0002-5507-0300.

(Dentith 2018; Fallis and Mathiesen 2019; Gelfert 2018; Jaster and Lanius 2018; Mukerji 2018), and on countering its apparently pernicious effects (Croce and Piazza 2021a; Fritts and Cabrera 2022; Rini 2017; Wright 2021). This paper won't offer yet another definition of fake news in general. Instead, it aims to identify a kind of fake news that is not captured by most accounts. This fake news is not intended to deceive those who consume it, neither about its content nor about the intentions of those who produce it. I thereby side with those (few) accounts of fake news that drop a deception condition (principally, the account developed by Pepp, Michaelson, and Sterken 2019), though I do not attempt to assess whether these accounts adequately capture the full range of fake news.

Those few philosophers who join with me in recognizing that fake news need not seek to deceive have proposed *inward*-looking explanations of its function: they maintain that such fake news is consumed and shared due to the roles it plays in exchanges within an in-group. I aim to supplement, not supplant, such explanations: I find at least some of them highly plausible. I will argue that in addition to these inward-looking functions, fake news also plays an *outward*-looking role: I will suggest that fake news that is not intended to deceive its consumers (from now on, *non-deceptive fake news*) is produced and consumed, in important part, because of the role it plays as evidence in arguments with out-group members. Of course, it's bad evidence: it is, after all, it is *fake* news. But it needs to possess only some degree of surface plausibility, not genuine credibility, to provide citable evidence for political positions. I will suggest that understanding this outward-looking function that fake news is often designed to play, and does play, is necessary for addressing all the epistemic harms it gives rise to.

Inevitably, my account of the function of non-deceptive fake news is speculative. It is intended to offer a plausible explanation for a phenomenon that I take to require one, and that is not accounted for by extant accounts such as those offered by Marianna Ganapini (2021) or Michel Croce and Tommaso Piazza (2021a). Because these accounts are inward-looking, they explain intragroup behavior. But fake news also features in outward-looking behavior, and it is this behavior I aim to explain. [1]

[1] It's worth noting that I do not take the categorization of fake news into deceptive and non-deceptive to be exclusive and exhaustive. One and the same story may deceive some and not others who nevertheless consume it and repeat it (indeed, the widespread repetition of fake news by non-believers may lead to some consumers believing it). I'm grateful to a reviewer for this journal for highlighting this point.

I. Fake News and Deception

Perhaps guided by the word 'fake', philosophers have almost universally identified fake news with apparent news that is deceptive. On most accounts, fake news is produced or disseminated with the intention to deceive the audience (Dentith 2018; Gelfert 2018; Rini 2017). On these accounts, fake news consists, roughly, in information presented to look like genuine news but which is false, and is aimed at instilling false belief. Opposition to the deception condition has come largely from those who object that fake news can be bullshit, in Frankfurt's sense: that is, produced by people who are indifferent to the truth of what they're saying, and who therefore lack the intention to deceive (Frankfurt 2009). Mukerji *defines* fake news as bullshit asserted in the form of a news publication (Mukerji 2018). However, though proponents of the bullshit account recognize that producers of fake news need not seek to deceive their audience *about the content of the story*, they do not drop the deception condition entirely (Grundmann 2023). Mukerji claims that the producer must attempt to deceive their audience *about their motives*. Duncan Pritchard has recently gone further, arguing that even if the producer's primary intention is to make money and not to deceive the audience, they will need to make their stories plausible enough for people to click on them, and must therefore be concerned with deceiving the audience (Pritchard 2021).

There is little doubt that some people believe some fake news (that is, they take the content conveyed by what I will call the story to be true, or probably true). Nor is there much doubt that some fake news is produced with the intention to deceive. The best evidence we have that some people believe some fake news is that they go on to act consistently with it, in some circumstances in which they've staked something significant on its being true. Pizzagate provides a compelling example. According to the Pizzagate story, senior members of the Democrat party were running a child sex ring out of the basement of a D.C. Pizza restaurant. Notoriously, one consumer of the rumor was sufficiently convinced by it that he drove from North Carolina to D.C. to investigate. Edgar Maddison Welch stormed the restaurant armed with an assault rifle, firing several shots before being arrested. Welch is not the only person sufficiently convinced by fake news to act on it in a high-stakes situation. Tragically, hundreds of Iranians were sufficiently convinced by the rumor that alcohol can prevent Covid-19 to consume an illegal homebrew, with many subsequent deaths (Associated Press 2020).

There's also plenty of evidence that producers of fake news sometime aim to deceive their audience. For instance, some fake news seems to have been

produced with the aim of shifting market prices (Merle 2017). There have also been multiple attempts to spread rumors, sometimes in the form of apparent news, with the aim of influencing election outcomes. For example, fabricated tweets aimed at influencing votes circulated widely during the last Australian federal election (Jensen 2019). Fake news is often designed to deceive, and some people are deceived. It's no part of my view to deny these facts.

But we should resist thinking that fake news, even widely circulated fake news, succeeds in deceiving nearly everyone who shares it, 'likes' it or of those who endorse it or repeat it (Croce and Piazza 2021b; Ganapini 2021). Often, it doesn't even deceive a majority of these people. One reason to doubt that fake news is always believed by most of the people who respond favorably to it is its often incredible nature and content. [2] Pizzagate was based on the flimsiest of fabricated evidence: an implausibly strained decoding of the DNC emails released by Wikileaks (with 'cheese pizza' alleged to mean 'child pornography', simply on the basis of the fact that they share the same initial letters). The Wayfair conspiracy theory was based on little more than the fact that some internet users found the price of a cabinet the furniture retailer was selling exorbitant (Funke 2020). Some people are gullible and no doubt there are true believers in all of these conspiracy theories. But when incredible and obviously baseless rumors go viral, we should resist the temptation to think that a large proportion of our fellow citizens are so gullible.

In fact, most of the people who repeat and appear to accept such rumors do not go on to act consistently with them, even when they could at low cost to themselves. Hugo Mercier gives the example of the anti-Semitic rumor that swept the town of Orleans in 1969, that Jewish shopkeepers were kidnapping and selling local girls (Mercier 2020). At the height of the panic, those who claimed to accept the rumor went so far as to stare hard at the offending stores; they didn't raid them or even demand police action. For every Welch willing to 'self-investigate' fake news with a gun, there are thousands and thousands of people who share it and cite it in political discussion, but don't take it seriously enough to act on it.

[2] We should not too hastily infer a lack of genuine belief from apparent bizarreness of content. What strikes an agent as incredible will vary from person to person, depending on their background beliefs. Given the right background beliefs, almost any claim can be rationally accepted (Levy 2021). A bizarreness criterion is more easily applied *within* a particular culture than *across* cultures: some claims are incredible for almost every person within a culture, insofar as cultures are (partially) constituted by background beliefs. Some of the claims that conspiracy theories assert are surely incredible by *their own lights*.

There's also experimental evidence that more people express support for a range of partisan views than genuinely believe them. One well-known study focused on a notorious episode from the earliest days of the Trump presidency. In January 2017, Trump's then press secretary Sean Spicer stated that the crowd at Trump's inauguration had been the biggest in the history of the event. Photographic evidence and mass transit records provided conclusive evidence that Spicer was wrong. Despite the evidence, Kellyanne Conway defended what she called Spicer's "alternative facts." Her defense of Spicer, more than Spicer's claims themselves, caused a media storm and a great deal of consternation over the possibility we're living in a post-truth era.

Whether Spicer and Conway's assertions, or coverage of them in Trump-friendly media, constituted fake news is not a question I will attempt to answer. What is more important to me is the attitude to the controversy taken by Trump supporters, insofar as evidence concerning their attitudes is likely to generalize to more paradigmatic fake news. Evidence for their attitudes comes from a study by Schaffner and Luks (Schaffner and Luks 2018; see Ross and Levy 2022 for replication). While the controversy was still fresh, they gave participants in their studies unlabeled photographs depicting the 2009 Obama inauguration and the 2017 Trump inauguration, and simply asked them to report which depicted a bigger crowd. The results were striking. In defiance of the clear photographic evidence, 15% of Trump voters (but only 3 and 2% of non-voters and Clinton voters, respectively) chose the photo of the Trump inauguration as depicting a bigger crowd.

By far the most plausible explanation of this difference is that Trump supporters were engaged in *expressive responding* (sometimes called 'partisan cheerleading'). They didn't report a sincerely-held belief: they expressed support for their side of politics. Further evidence for this claim is that better educated Trump voters were *more* likely than the less-well educated to choose the 2017 photo. Presumably, better educated voters were more likely to recognize the photos as depicting the Trump and Obama inaugurations, and to recognize an opportunity to express their support for him by making their choice (the few Clinton voters and non-voters who chose the Trump photo were instead probably engaged in a kind of trolling of the experimenters (Lopez and Hillygus 2018).

Expressive responding appears to be common, though its scope and the size of the effect are controversial. A broad range of experimental and survey evidence appears to indicate that people often report attitudes they don't sincerely hold (Hannon 2021; Levy and Ross 2021). For example, US respondents tend to report that the economy is in a worse state when the president be-

longs to a rival political party than when the president belongs to the party they support, but their economic decisions seem to better reflect the actual state of the economy than their reported assessment (Bullock and Lenz 2019). On the other hand, some experimental work has failed to find evidence of expressive responding, even when responding accurately is incentivized (Berinsky 2018). There are at least two possible ways to explain these conflicting data. It might simply be that it is difficult to incentivize accuracy. Or the conflicting data might indicate that the prevalence of expressive responding varies very significantly from topic to topic, in ways that diverge from what we might have expected (Graham and Yair 2022). Whatever the explanation, it seems overwhelmingly likely that many political partisans do not believe some of the fake news they nevertheless consume, share and verbally endorse.

If the scope and the degree of expressive responding varies considerably from topic to topic, measuring the size of the effect will be extremely difficult. Prior et al. found that (relatively) small monetary rewards for correct responses reduced partisan bias from 12 percent to 6 percent (Prior, Sood, and Khanna 2015). It is likely that on many topics, the size of the effect is much bigger. Note, first, that financial incentives may provide perverse incentives. If reporting that *I believe the economy is bad* allows me to signal my opposition to the current president, the opportunity to turn down an incentive to report the economy is doing well offers me the opportunity to send a stronger signal. Second, I may see my support of my side as politics as a sacred value (Tetlock 2003); people are strongly resistant to betraying sacred values for money. Third, it is likely that the kinds of topics Prior et al. probed – perceptions of the state of the economy – do not rouse polarized passions to the same degree as culture war issues. Estimating the true size of the effect is currently impossible, given the available experimental evidence, but there is little doubt at least a substantial minority of people sometimes report things they don't really believe to express support for 'their' side. Another, unknown, proportion of people report things they don't believe to 'troll' the experimenters or for a prank (Lopez and Hillygus 2018). Scott Alexander has proposed a 'lizardman constant': the allegedly constant rate at which people will insincerely respond 'yes' to the question "are lizardmen running the earth?" (Alexander 2013). I am skeptical that there is any such constant; I suspect that the proportion of insincere responses is highly context-dependent (how willing are people to participate in the survey, which is a function of sample selection and incentives, amongst other things; the content of the questions, with more bizarre content selecting for an increased irritation with them; the length of the survey; the degree to which the sample's political and social views depart from

those they attribute to the experimenters, and so on). While the degree of expressive responding and sheer trolling surely varies from context to context, insincere responding probably accounts for some proportion of answers on most surveys.

There is every reason, therefore, to think that many people who report accepting a conspiracy theory or a rumor (especially a bizarre one) do not genuinely believe it. Given this fact, I will suggest, we should look for other functions of fake news to explain its production and consumption, and the way in which it is put to use in political argument.

II. Non-Doxastic Explanations of Fake News

If fake news need not be believed by those who consume, like, and share it, what explains its production and consumption? There are two existing accounts of the function of (some) fake news that aim to explain why it might be produced and consumed in the absence of an intent to deceive and of actual deception. Marianna Ganapini suggests that fake stories need not be believed, or even believable, in order to play a *signaling* function (Ganapini 2021). [3] As she argues, the absurdity of fake news can be a feature, rather than a bug, when it comes to playing this role. A willingness to believe, or pretend to believe, bizarre claims is a reliable signal of commitment to an in-group, because it exposes one to ridicule and risks social ties with the out-group (Mercier 2020). As Ganapini shows, fake news can signal commitment to the group, enable the coordination of group action and solidify group identity: for none of these roles is it essential that the fake news be believed or believable. [4]

[3] Ganapini explicitly focuses on fake stories, rather than fake news, because she takes the concerns about that term offered by philosophers such as David Coady (2021) and Joshua Habgood-Coote (2019; 2020) seriously. I am less moved by these worries (I discuss them briefly in note 7). While her category encompasses some stories that are not plausibly regarded as fake news and excludes some that are, her account is designed to explain much of what other people call fake news. Since my account depends on fake stories having the trappings of news, or other cues for epistemic respectability, it does not aim to explain the function of fake stories that lack these properties.

[4] Funkhouser (2017) argues that some belief reports have the function of signaling commitments to other agents. But on his account, the signals are not themselves beliefs and need not entail them. Signaling requires that we represent ourselves as believing, whether or not we genuinely believe in the content asserted. The signal I send by asserting *the election was rigged* is reliable, not because it indicates that I believe that the election was rigged but because it indicates that I am committed to our side.

In more recent work, Ganapini (2022) has offered an even more inward-looking account of the role played by fake stories. She argues that even when their content is not believed to be literally true, their narrative structure nevertheless allows those who consume and share them to take them to be credible; that is, as providing insight into the kind of events that could occur. The Democrats might not have literally engaged in satanic rituals, but the plausibility of the story reveals their moral depravity (more recently Langdon et al. 2024 advanced a similar account). Ganapini argues that fake stories that are not believed can play an identity-protective role, shielding consumers from counterevidence and easing cognitive dissonance. Whereas her signaling account focuses on the role that fake stories play within the group, her (compatible) narrative account focuses on the role they play within the *individual*. That's as inward-looking an account as can be imagined.

Michel Croce and Tommaso Piazza's account of non-deceptive fake news is also inward-looking (Croce and Piazza 2021b). Croce and Piazza argue that the producers of fake news need not intend to deceive: rather, they may be unconcerned with the truth at all. Such producers thereby exhibit the vice that Quassim Cassam (2018) calls epistemic insouciance (though Croce and Piazza do not use that term). Producers need not be concerned with even an appearance of truth, because consumers in turn do not seek to be informed or to inform, but instead seek what Croce and Piazza call 'social recognition': to build up ties of solidarity with one another and to feel recognized as part of a community. It is enough that the story has a content that group members would *like* to be true for fake news to play these roles. It need not be true, nor taken to be true.

More recently, Dan Williams (2023) has proposed an account of the function of rationalizations that is designed to encompass at least some fake news. On Williams' account, there is a market for rationalizations of unfounded beliefs, driven by consumers' desire to believe congenial claims. The overlap between the cases my account aims to explain and those Williams aims to explain may be quite small: while rationalizations are consumed only because truly compelling evidence isn't available, consumption aims at genuine belief and successful rationalizations will therefore be reasonably plausible, at least to the casual eye. Williams therefore seems to exclude more incredible fake news from the scope of his account (in fact, his main target is not, quite, fake news at all, but opinion offered by partisan hacks).

A more important difference between Williams' account and the one I offer is that his is (once again) inwards-looking. Williams is explicitly concerned with the utility of beliefs for individuals: rationalizations have an i*ntra*psychic

function. Like Ganapini's more recent account, it is as inward-looking as can be: concerned with the psychological states of individuals, not the group and certainly not outsiders.

All these accounts explain the production and consumption of non-deceptive fake news in ways that turn on its function within a group of like-minded people. On Ganapini's accounts, fake news allows group members to signal to one another and to protect the identity of individual group members. On Croce and Piazza's account, fake news is shared within the group to tighten bonds of solidarity. On Williams account, (some) fake news is consumed for the purpose of individual belief-based utility. None of these accounts adverts to an outward-facing role.

Fake news surely plays multiple roles, and the cases presented for these accounts are persuasive. Each likely captures some of the reasons people share and consume fake news. But, I will argue, non-deceptive fake news also has a surprising *outward* facing role. Non-deceptive fake news consists in claims that (a) are minimally defensible, in virtue of their content or (more usually) in virtue of being attributable to an epistemic authority and (b) therefore can be cited in political argument and in maintaining political commitments. Such fake news need not be, and often is not, believed by those for whom it serves these functions.

III. Fake News as a Rhetorical Weapon

Let's begin with (b); the role that such fake news plays in arguments. Non-deceptive fake news provides a resource to political partisans: it can help them to *win* arguments. There is of course a sense of 'winning an argument' in which only good evidence is relevant to which argument wins. It's that sense that is paramount in philosophical debates. But in addition to this truth-centered sense of 'winning an argument', there's a purely pragmatic sense. In this sense, winning an argument doesn't require actually having better evidence than opponents, nor does it require convincing them or even convincing the antecedently agnostic. Winning an argument in this pragmatic sense consists in carrying the day (advancing one's agenda, seeing off criticism), and one can win arguments in this sense even when the evidence one cites in favor of one's views is mainly bluster and even when it's seen to be bluster. For example, a politician accused of spending public money to bolster her re-election chances, rather than to advance the public good, might win the argument, in *this* sense, by giving a response that seems sufficiently

plausible and sufficiently relevant to withstand public scrutiny long enough for the fuss to die down.

Real life examples are easy to find; most interviews with politicians seem to feature at least one attempt to win arguments using transparent bluster. A recent case involves the now-former Australian prime minister, Scott Morrison, following revelations that spending announcements made just prior to the last election targeted seats his party was in danger of losing (rather than following the politically neutral, needs-based, process that was supposed to be used). Asked whether he was okay with such spending being directed in a blatantly partisan way, Morrison replied "I'm very OK with the idea of building car parks to ensure people can get a park, get on a train, can get to work sooner, get home sooner, because urban congestion and people commuting is a daily challenge" (Murphy and Karp 2021). Morrison's rhetorical move here is a familiar one: rather than directly responding to the question put to him, he reframes it and addresses a related question. He puts claims to use for rhetorical purpose, to defend his government's actions. As far as I can gauge, Morrison won the argument: journalists couldn't accuse him of entirely dodging the question and eventually moved on to other matters.

While Morrison's argumentative strategy didn't invoke fake news, his argumentative strategy nicely illustrates the kind of use to which fake news may be put. He didn't genuinely report a mental state – not, at least, one that he sincerely took to be directly relevant to the reporter's question – he wielded words as rhetorical weapons. Their purpose wasn't to respond to the question, but to see it off. To this end, politicians and ordinary people may cite whatever comes handy. If it's true, so much the better. But it need not be true: it need only be presented as taken to be true and somewhat plausible to serve its function in argument. Think of Trump's suggestion that Ted Cruz's father was somehow involved in the Kennedy assassination. When Cruz was a rival for the Republican nomination, Trump used the insinuations to good rhetorical effect. He won the argument, at least by the principal measure that mattered to him (and to Cruz): he secured the nomination.

At least prior to Trump, politicians generally refrained from citing blatantly fake news in arguments, most likely because doing so might damage their credibility. Ordinary people may do so much more freely, because they don't have to be as concerned with the perceptions of more neutral observers. Facebook and Twitter arguments are routinely won, in the pragmatic sense – the conversation is allowed to move on – via the citation of fake news.

Non-deceptive fake news also plays a different, inward-facing, role: it allows group members to maintain their commitment to their views. We are, af-

ter all, reason giving animals: it is important to us that we are able to cite evidence in favor of our views. Fake news may provide us with such evidence. Most of us are at least somewhat sensitive to the quality of evidence, and we prefer to rely on genuinely strong evidence. But we're under rational pressure, from ourselves and from others, to say *something* in response to challenges to our views, and we will respond with the best evidence we have. If we have nothing better that is sufficiently relevant than fake news, many of us may cite that (those who are more epistemically conscientious may refrain from citing obviously fake news, but many of us might nevertheless frantically google in response to social media challenge and cite the first apparently plausible source we can find). We thereby satisfy ourselves that there are good reasons for our views, and we relieve ourselves of the need to revise them.

Being citable in argument and enabling those who share and consume it to maintain their commitments is a central function of fake news, I suggest, one that helps to explain why it is widely produced, consumed, and shared, even when it is incredible, and in the face of the evidence that sometimes at least, few of the people who endorse it are genuinely convinced by it. Claims wielded as argumentative weapons need be seen as sufficiently relevant to allow the conversation to move on or for partisans to be satisfied with their views.

What makes non-deceptive fake news plausible enough to play these roles? Most political facts are known to most of us via testimony, not direct observation, and in most cases, plausibility will stem from some combination of content ("that's the kind of thing a Democrat *would* do") and trust in the source of the testimony. It is in virtue of attributability to what I'm calling an epistemic authority that such fake news is seen as sufficiently plausible. An epistemic authority is a person or institution widely recognized as a reliable source of information on a particular topic. Scientists are epistemic authorities regarding their specialty; athletes are epistemic authorities about their sport, and so on. Claims are more plausible in virtue of being attributed to an epistemic authority. Fake news is paradigmatically presented as stemming from a domain-general epistemic authority: the media. [5] Paradigmatic fake news adopts the

[5] Coady (2021) and Habgood-Coote (2019) both worry that talk of fake news serves a propaganda function by causing us to overlook how unreliable and tendentious the 'mainstream media' often is. It's worth emphasizing here that my account of a central function of fake news is neutral on whether what we call fake news is really very much less reliable than the legacy media. My account requires that (some of) the media is seen as an epistemic authority, not that that perception is accurate.

trappings of journalism, and it is in virtue of this fact that it can be attributed, with sufficient plausibility, to an epistemic authority.

The trappings of journalism confer epistemic respectability because they are evidence that a story has been generated in an epistemically conscientious manner: by people who care whether it is true and have employed methods that substantially raise the probability that is true (using sources that can be verified, double checking, consulting experts, and so on). Most of us may not know much about how news is generated, but we know that is supposed to use methods that make it reliable, and the format and characteristic language of news is a cue that such methods were employed. Fake news can with sufficient plausibility be attributed to an epistemic authority because it comes with the trappings of journalism.

Claims that play the rhetorical roles often played by fake news need not be news at all (nor need they be fake). What is needed is that the story can be attributed with significant plausibility to an epistemic authority, and such (genuine or apparent) authorities extend well beyond the media. It is fake news that paradigmatically plays this rhetorical role, because it is the media that is the main source of our political information (and politics is the arena in which we're most apt to rely on claims to shore up our sense of ourselves and to engage in arguments). But a great many other sources of information are sometimes politically relevant. We might rely on a (real or fake) scientific article about the efficacy of gun control. We might rely on purported eyewitness testimony about the behavior of police at a political protest. We might rely on a YouTube video by a supposed historian concerning the real causes of the Civil War. In each case, we cite apparent evidence that can with sufficient plausibility be attributed to an epistemic authority. So much the better if the apparent authority is a genuine authority.

Of course, the cues in virtue of which we attribute a story to an epistemic authority (e.g., the trappings of journalism) are often genuinely evidence of reliability. Some consumers of fake news, those who are somewhat convinced by it, may take these cues as evidence of genuine reliability. For many, however, they are not regarded as evidence; not, at least, as evidence worth putting real weight on. For these consumers of fake news, possession of the trappings of journalism may be enough to rely on the story in argumentative and signaling contexts: these trappings are not good evidence that the contents are true, but instead are cues in virtue of which the contents can be attributed to a putative epistemic authority. Even these consumers will generally be well aware that some sources are better than others, and that journalistic sources and practices differ in epistemically relevant way. Most of us, even those of us who

will resort to fake news when it's the best option available to us, would prefer to rely on more reliable sources if we can. We'd cite the *New York Times* or *Fox* if we could, but if we can't, we might cite *The Denver Guardian*. We'd prefer to cite a highly plausible story, but if we can't, we'll cite one about a Democrat child trafficking ring.

Those who rely on low quality and fake sources even when they don't take their possession of the trappings of journalism or similar such cues to provide genuinely good evidence need not be completely cynical. Many people will rely on such sources, when nothing that seems better is available, only when they're not very obviously false, by the person's own lights. There is a penumbra of truth, as it were, in which many such stories might fall. Running a child trafficking ring seems like it could just be the kind of thing that Democrats might do, and if they're not doing that, they're probably doing something pretty bad, so why not cite this story? We don't recognize ourselves as cynically relying on the fake; rather, we don't look too hard at our sources when we suspect they might not bear close scrutiny. [6]

As mentioned at the outset, my account is not unique in dropping a deception condition. Pepp, Michaelson and Sterken (2019) argue that a story becomes fake news when it is widely spread by people who *treat* it as having been produced by 'standard journalistic practices' though it was not. They recognize citing a story in a political argument as one way in which we can treat it as having been produced by such practices, and they require neither that those who treat a story in this way believe it nor that those who produce fake news intend to deceive. They also recognize how the trappings of journalism (and, on my view, other kinds of markers for epistemic authority) inspire trust, even in those who are not really deceived by the fake news. [7] How-

[6] Van Leeuwen (2014; 2023) argues that religious and ideological credences are not genuine beliefs; rather they are "secondary cognitive attitudes" somewhat akin to imaginings. We might wonder whether reliance on fake news also involves such secondary attitudes, especially given that his account aims to explain the same sort of divergence between asserted belief and behavior I've cited. The experimental evidence that people do not take themselves to believe the contents they assert suggests, however, that we can explain the divergences in these cases without postulating any novel attitudes: people are willing to assert claims when they're conscious they don't have to place any significant weight on them to express support for their side of politics. It remains possible that Van Leeuwen's account offers the best explanation of the cases he is centrally concerned with.

[7] One reason to regiment our use of 'fake news' to pick out only those instances which assume the trappings of journalism, and not other markers of reliability, is that such instances seem most resistant to the criticism that the term is redundant (Coady 2021; Habgood-Coote 2019). In his response to Brown (2019) and Pepp, Michaelson, and Sterken (2022), Habgood-

ever, their reasons for rejecting the deception condition are very different from mine, because they have not recognized the function of non-deceptive fake news.

Pepp, Michaelson and Sterken reject the deception condition on two grounds: first, producers of fake news may not aim to deceive because they actually believe the fantasies they promote; and second because their motivation might be money, not deception (they note that fake news could be produced by a bot: an algorithm incapable of possessing intentions might learn to produce apparent news as a means of generating click through). All that seems correct, but it overlooks a central reason for the production of non-deceptive fake news, a reason that stems from the function that such fake news plays in political argument. Producers may intentionally produce stories that they regard as only superficially plausible to serve as rhetorical weapons for their audience. Producers of fake news may generate it in order for their audience to cite it or otherwise rely on it in signaling and argument, as well as for social recognition. Alternatively, they may generate it for their own use: so the *producer* may cite it or otherwise rely on it. None of these motives require an intention to deceive on the part of the producer, nor does it require that consumers are deceived.

Nothing in my account of non-deceptive fake news is inconsistent with the account developed by Pepp, Michaelson and Sterken. It goes further than theirs inasmuch as it identifies the functions of such fake news, and in light of that function offers (what I take to be) a satisfying explanation for why fake news need not be produced with the intention to deceive. Producers and consumers may collaborate in the production of rhetorical weapons. I take this point to be an important one, inasmuch as it will provide us with tools for understanding the motivations of and the relations between prominent producers of fake news (and those who signal boost it without – quite – endorsing it, like Fox News) and their audiences.

Coote (2020) concedes that the case for a neologism is strongest with regard to false or misleading stories that mimic genuine news. I am not much moved by Habgood-Coote and Coady's argument because (a) I don't share the sense they have that the semantics of 'fake news' is particularly unsettled and (b) I agree with Brown and with Pepp, Michaelson and Sterken that the political risks of fake news are just as acute with many other political terms, including the terms that Habgood-Coote and Coady urge as replacements for 'fake news' (propaganda, and the like).

IV. The Harms of Fake News

If the forgoing is correct, then some fake news is produced without any intention to deceive and in fact does not deceive much of its audience (one and the same story can deceive some audience members, of course, and not others, though both groups rely on it in signaling and argument). Not all fake news fits this model, and not only fake news fits this model either: many things beside news can be relied on for argument and signaling, in the absence of a belief. But if fake news often does not deceive anyone, what epistemic harm does it do? Knowledge and belief are, of course, the central concepts of epistemology, so the widespread assumption that fake news is harmful because it produces false belief is natural. But fake news can be very harmful even when no one is deceived. I focus exclusively on those harms that arise from the outward-looking uses of fake news.

The central outward-looking use of fake news, I have argued, is as a weapon in political arguments. At its best, political debate has a number of positive contributions to make to our lives as citizens. Its major role is epistemic: political debate is a contest of ideas, and allows those engaged in the debate, as well as third-parties, to learn about these ideas, and to change their minds when appropriate. While we are often cynical that arguments really change minds about politically charged topics, there is evidence that they have some, albeit limited, effect on most of us (Coppock 2023). Political debate also has non-epistemic roles to play. It can promote political engagement, which is important to the functioning of democratic institutions.

When fake news is wielded as a weapon in the sort of way I've described, it is antithetical to these roles. Since it is often quite transparently fake, it promotes cynicism about politics, rather than engagement. Insofar as it is successful, it dissociates argumentative success from actual evidence, thereby ensuring that the opportunity for participants and onlookers to learn is lost. It turns debate into bad theatre. When it comes to circulate widely because the norms tolerate reliance on it, we may easily conclude that news is not worth consuming at all. [8] Blake-Turner has recently argued that fake news degrades our epistemic environment and undermines epistemic confidence (Blake-Turner 2020). His focus is on genuinely believed stories, but similar considerations apply to fake news produced and wielded as a rhetorical weapon. We may not come to believe fake news, but we may find it prohibitively difficult to come

[8] Bernecker (2021) defends such abstinence.

to well justified beliefs at all in an environment in which fake news dominates so much discussion.

In this sort of way, non-believed fake news might have doxastic harms. If fake news on a particular topic circulates widely, then we may have trouble discerning the truth about it. Steve Bannon's strategy of "flooding the zone" might be understood along these lines (Illing 2020). Bannon appears to aim not to convince people of anything, but to make it difficult for them to from beliefs on the topic. This kind of strategy is allegedly the hallmark of Russian propaganda; again, the aim is not to convince but to undermine the capacity to form confident beliefs at all (Pomerantsev 2014). [9]

Non-deceptive fake news may also be an obstacle to public debate and political compromise or progress. It allows partisans to continue to fight battles that they have obviously already lost, so far as the merits of their position are concerned. It therefore is an obstacle to moving on. Suppose that few politicians actually accept that Trump really won the 202 election. Nevertheless, their continued public advocacy for this view, and the ways in which it allows them to signal to their base and to win, in the pragmatic sense, political arguments, ensured a lack of felt pressure to work with the Biden government. They could continue to fight the last battle instead. There might be far more agreement about global warming than is commonly thought (the finding that better educated Republicans are more, not less, likely to reject it (Kahan 2015) suggests that rejection might be more verbal than substantive), but so long as a large class of people feel able to assert its non-reality, they can block effective action.

Finally, the widespread invocation of fake news might lead to polarization. As Michael Hannon shows, evidence for polarization in the political beliefs of partisans is not strong, but there is good evidence of *affective* polarization (Hannon 2021). We may not diverge in our beliefs about policy more now than in the past, but we think much less well of political opponents. Reliance on fake news could exacerbate affective polarization, because such reliance requires representing oneself as believing it. Far fewer people genuinely believe the QAnon conspiracy theory than share it or verbally endorse it, but its broad spread and enthusiastic embrace leads many to think that a large proportion of their fellow citizens have, in the words of one pollster, gone "bonkers"

[9] I am grateful to a reviewer for this journal for highlighting the doxastic effects of unbelieved fake news. The reviewer also notes that the prevalence of fake news may lead those who argue in good faith to withdraw from the public sphere (or alternatively, to become difficult for others to identify), so that again its widespread circulation may undermine belief.

(Rose 2020). [10] In this climate, we're less likely to see political opponents as reasonable people who might be worth trying to persuade.

How should we address these harms? This is a harder task, at least in some ways, than it would be were those people who share and endorse fake news always deceived by it. If that were the case, we might hope to address the problem via the solution advocated by Lee McIntyre: teaching critical thinking in schools (McIntyre 2018). Teaching such skills would have a limited effect, I predict. Far fewer people than is commonly thought genuinely believe implausible fake news, and I suspect those few who do are unlikely to be good candidates for substantial gains from critical thinking. Many more people are taken in by plausible looking stories, of course, but it's far from clear that critical thinking would allow them to identify such stories: stories that look plausible may have the kinds of features that pass such scrutiny. There is of course a continuum from credible but false to incredible, and we should expect critical thinking skills to get traction in some cases and for some people. At very best, though, better critical thinking will fall short of dealing with all the harms that arise from fake news.

If we are to address the problem of non-deceptive fake news, we might do better to focus on the social norms that tolerate its propagation, sharing and even reliance on it in argument. Shifting such norms is extremely difficult, especially given the fact that most of us are disposed to do something similar (if not to rely on the fake, at least to cite the apparently credible without looking too hard at it), and the fact that we tend to give 'our' side of politics a pass when they cite the unreliable. Perhaps the introduction and enforcement of a norm of calling out those who engage in these practices could help to turn the tide. Since political opponents can be expected to dismiss such calls as having a partisan motive, we'd do better to focus on our own side. Such norm shifting might be assisted by more responsible reporting about fake news. Right now, even those media outlets that aim for accuracy contribute to the impression that a very large proportion of people believe obviously fake news. In doing so, they might inadvertently make it easier to rely on it. Wider publicization of

[10] The pollster was commenting on an NPR/Ipsos poll that found that 17% of Americans accept the QAnon conspiracy, with another 37% being unsure whether it is true (Ipsos 2020) But polls like this transparently probe political identity and many people respond accordingly (QAnon support was unsurprisingly driven by Republicans). Moreover, the poll had design flaws known to increase the likelihood of expressive responding. It gave people true/false/don't know as options. People are reluctant to report ignorance, and the provision of a 'skip' option, which allows those who don't know the right answer to a question to continue without confessing ignorance, reduces guessing and partisan response (Motta et al. 2019).

the fact that many of those who endorse fake news don't really believe the claims they make might undermine their credibility in relying on it. These suggestions are of course tentative and sketchy. They await elaboration (or replacement by better proposals) by those with the skills to do better than I can, guided, I hope, by my account of the function that non-deceptive fake news plays. [11]

Conclusion

Contrary to the orthodoxy in philosophy, fake news need be neither deceptive nor deceiving. People routinely rely upon, often by endorsing, claims they don't believe in order to express support for their side of politics. There's every reason to think they share, like and endorse fake news for the same purposes. Very likely, they also thereby signal group membership to fellow partisans. They also sometimes rely on fake news to win political arguments. Though some fake news is intended to deceive, and some people are deceived even by bizarre fake news, much circulates widely without deception.

The good news is that our fellow citizens are much less credulous than we commonly think. The bad news is that fake news that isn't believed may be every bit as harmful as fake news that is believed. It's a very significant epistemic pollutant and reducing its prevalence would be valuable. Censorship and control by government or multinational corporations have their obvious risks, of course. I've suggested that we target the social norms that tolerate reliance on it instead. We shouldn't be very confident that such an approach will succeed, but given the costs of fake news, and the minimal risks of this response, it is surely worth the attempt.

Works Cited

Alexander, Scott. 2013. 'Lizardman's Constant Is 4%'. *Slate Star Codex* (blog). 12 April 2013. https://slatestarcodex.com

Associated Press. 2020. 'In Iran, False Belief a Poison Fights Coronavirus Kills Hundreds'. Snopes.Com. 27 March 2020. https://www.snopes.com

[11] See Westra (2021) for a defense of the claim that changing our behavior online can contribute to changing social norms, and for better developed proposals for how we might go about this.

Berinsky, Adam J. 2018. 'Telling the Truth about Believing the Lies? Evidence for the Limited Prevalence of Expressive Survey Responding'. *The Journal of Politics* 80 (1): 211-24. DOI: 10.1086/694258.
Bernecker, Sven. 2021. 'An Epistemic Defense of News Abstinence'. In *The Epistemology of Fake News*, edited by Sven Bernecker, Amy K. Flowerree, and Thomas Grundmann, 286-309. Oxford University Press.
Blake-Turner, Christopher. 2020. 'Fake News, Relevant Alternatives, and the Degradation of Our Epistemic Environment'. *Inquiry: An Interdisciplinary Journal of Philosophy*. DOI: 10.1080/0020174X.2020.1725623.
Brown, Etienne. 2019. '"Fake News" and Conceptual Ethics'. *Journal of Ethics and Social Philosophy* 16 (2). DOI: 10.26556/jesp.v16i2.648.
Bullock, John G., and Gabriel Lenz. 2019. 'Partisan Bias in Surveys'. *Annual Review of Political Science* 22 (1): 325-42. DOI: 10.1146/annurev-polisci-051117-050904.
Cassam, Quassim. 2018. *Vices of the Mind: From the Intellectual to the Political*. Oxford University Press.
Coady, David. 2021. 'The Fake News about Fake News'. In *The Epistemology of Fake News*, edited by Sven Bernecker, Amy K. Flowerree, and Thomas Grundmann, 68-81. Oxford University Press.
Coppock, Alexander. 2023. *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago Studies in American Politics. Chicago, IL: University of Chicago Press.
Croce, Michel, and Tommaso Piazza. 2021a. 'Consuming Fake News: Can We Do Any Better?' *Social Epistemology* 0 (0): 1-10. DOI: 10.1080/02691728.2021.1949643.
———. 2021b. 'Misinformation and Intentional Deception: A Novel Account of Fake News'. In *Virtues, Democracy, and Online Media*, edited by Maria Silvia Vaccarezza and Nancy Snow. Routledge.
Dentith, M. R. X. 2018. 'What Is Fake News?' *University of Bucharest Review*, no. 2, 24-34.
Fallis, Don, and Kay Mathiesen. 2019. 'Fake News Is Counterfeit News'. *Inquiry* 0 (0): 1-20. DOI: 10.1080/0020174X.2019.1688179.
Frankfurt, Harry G. 2009. *On Bullshit*. Princeton University Press.
Fritts, Megan, and Frank Cabrera. 2022. 'Fake News and Epistemic Vice: Combating a Uniquely Noxious Market'. *Journal of the American Philosophical Association* 8 (3): 454-75.
Funke, Daniel. 2020. 'How the Wayfair Child Sex-Trafficking Conspiracy Theory Went Viral'. Polifact. 15 July 2020. https://www.politifact.com
Funkhouser, Eric. 2017. 'Beliefs as Signals: A New Function for Belief'. *Philosophical Psychology* 30 (6): 809-31. DOI: 10.1080/09515089.2017.1291929.
Ganapini, Marianna. 2021. 'The Signaling Function of Sharing Fake Stories'. *Mind and Language*.
———. 2022. 'Absurd Stories, Ideologies, and Motivated Cognition'. *Philosophical Topics* 50 (2): 21-39.

Gelfert, Axel. 2018. 'Fake News: A Definition'. *Informal Logic* 38 (1): 84-117. DOI: 10.22329/il.v38i1.5068.
Graham, Matthew H, and Omer Yair. 2022. 'Expressive Responding and Trump's Big Lie'. https://m-graham.com.
Grundmann, Thomas. 2023. 'Fake News: The Case for a Purely Consumer-Oriented Explication'. *Inquiry* 66 (10): 1758-72.
Habgood-Coote, Joshua. 2019. 'Stop Talking About Fake News!' *Inquiry: An Interdisciplinary Journal of Philosophy* 62 (9-10): 1033-65. DOI: 10.1080/0020174x.2018.1508363.
———. 2020. 'Fake News, Conceptual Engineering, and Linguistic Resistance: Reply to Pepp, Michaelson and Sterken, and Brown'. *Inquiry* 0 (0): 1-29. DOI: 10.1080/0020174X.2020.1758770.
Hannon, Michael. 2021. 'Disagreement or Badmouthing? The Role of Expressive Discourse in Politics'. In *Political Epistemology*, edited by Elizabeth Edenberg and Michael Hannon. Oxford University Press.
Illing, Sean. 2020. '"Flood the Zone with Shit": How Misinformation Overwhelmed Our Democracy'. *Vox*, 6 February 2020. https://www.vox.com
Ipsos. 2020. 'More than 1 in 3 Americans Believe a "Deep State" Is Working to Undermine Trump'. Ipsos. 30 December 2020. https://www.ipsos.com
Jaster, Romy, and David Lanius. 2018. 'What Is Fake News?' *Versus* 2 (127): 202-27.
Jensen, Michael. 2019. '"Fake News" Is Already Spreading Online in the Election Campaign – It's up to Us to Stop It'. *The Conversation* (blog). 24 April 2019. http://theconversation.com
Kahan, Dan M. 2015. 'Climate-Science Communication and the Measurement Problem'. *Political Psychology* 36 (s1): 1-43.
Langdon, Julia A., Beth Anne Helgason, Judy Qiu, and Daniel A. Effron. 2024. '"It's Not Literally True, But You Get the Gist:" How Nuanced Understandings of Truth Encourage People to Condone and Spread Misinformation'. *Current Opinion in Psychology* 57 (June):101788. DOI: 10.1016/j.copsyc.2024.101788.
Levy, Neil. 2021. *Bad Beliefs: Why They Happen to Good People*. Oxford: Oxford University Press.
Levy, Neil, and Robert M. Ross. 2021. 'The Cognitive Science of Fake News'. In *The Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen de Ridder, 181-91. Routledge.
Lopez, Jesse, and D. Sunshine Hillygus. 2018. 'Why So Serious?: Survey Trolls and Misinformation'. SSRN Scholarly Paper ID 3131087. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.3131087.
McIntyre, Lee. 2018. *Post-Truth*. MIT Press.
Mercier, Hugo. 2020. *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton: Princeton University Press.
Merle, Renae. 2017. 'Scheme Created Fake News Stories to Manipulate Stock Prices, SEC Alleges.' *Los Angeles Times*, 5 July 2017. https://www.latimes.com

Motta, Matthew, Daniel Chapman, Dominik Stecula, and Kathryn Haglin. 2019. 'An Experimental Examination of Measurement Disparities in Public Climate Change Beliefs'. *Climatic Change* 154 (1): 37-47. DOI: 10.1007/s10584-019-02406-9.

Mukerji, Nikil. 2018. 'What Is Fake News?' *Ergo: An Open Access Journal of Philosophy* 5:923-46. DOI: 10.3998/ergo.12405314.0005.035.

Murphy, Katherine, and Paul Karp. 2021. 'Australians Are the Winners': Scott Morrison Defends Controversial Commuter Car Parks Fund'. *The Guardian*, 5 August 2021. https://www.theguardian.com

Pepp, Jessica, Eliot Michaelson, and Rachel Sterken. 2022. 'Why We Should Keep Talking About Fake News'. *Inquiry: An Interdisciplinary Journal of Philosophy* 65 (4): 471-87. DOI: 10.1080/0020174X.2019.1685231.

Pepp, Jessica, Eliot Michaelson, and Rachel Katharine Sterken. 2019. 'What's New About Fake News?' *Journal of Ethics and Social Philosophy* 16 (2). DOI: 10.26556/jesp.v16i2.629.

Pomerantsev, Peter. 2014. 'How Vladimir Putin Is Revolutionizing Information Warfare'. *The Atlantic*, 9 September 2014. https://www.theatlantic.com

Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. 'You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions'. *Quarterly Journal of Political Science* 10 (4): 489–518. DOI: 10.1561/100.00014127.

Pritchard, Duncan. 2021. 'Good News, Bad News, Fake News'. In *The Epistemology of Fake News*, edited by Sven Bernecker, Amy K. Flowerree, and Thomas Grundmann, 46-67. Oxford University Press.

Rini, Regina. 2017. 'Fake News and Partisan Epistemology'. *Kennedy Institute of Ethics Journal* 27 (S2): 43-64. DOI: 10.1353/ken.2017.0025.

Rose, Joel. 2020. 'Even If It's "Bonkers," Poll Finds Many Believe QAnon And Other Conspiracy Theories'. *NPR*, 30 December 2020, sec. National. https://www.npr.org

Ross, Robert M., and Neil Levy. 2022. 'Expressive Responding in Support of Donald Trump: An Extended Replication of Schaffner and Luks (2018)'. PsyArXiv. DOI: 10.31234/osf.io/3fvyn.

Schaffner, Brian F., and Samantha Luks. 2018. 'Misinformation Or Expressive Responding? What An Inauguration Crowd Can Tell Us About The Source Of Political Misinformation In Surveys.' *Political Opinion Quarterly* 82 (1): 135-47.

Tetlock, Philip E. 2003. 'Thinking the Unthinkable: Sacred Values and Taboo Cognitions'. *Trends in Cognitive Sciences* 7 (7): 320-24. DOI: 10.1016/S1364-6613(03)00135-9.

Van Leeuwen, Neil. 2014. 'Religious Credence Is Not Factual Belief'. *Cognition* 133 (3): 698-715. DOI: 10.1016/j.cognition.2014.08.015.

———. 2023. *Religion as Make-Believe: A Theory of Belief, Imagination, and Group Identity*. Cambridge, MA: Harvard University Press.

Westra, Evan. 2021. 'Virtue Signaling and Moral Progress'. *Philosophy and Public Affairs* 49 (2): 156-78. DOI: 10.1111/papa.12187.

Williams, Daniel. 2023. 'The Marketplace of Rationalizations'. *Economics & Philosophy* 39 (1): 99-123. DOI: 10.1017/S0266267121000389.
Wright, Sarah. 2021. 'The Virtue of Epistemic Trustworthiness and Re-Posting on Social Media'. In *The Epistemology of Fake News*, edited by Sven Bernecker, Amy K. Flowerree, and Thomas Grundmann, 245-64. Oxford University Press.

DOI 10.17473/2240-7987-2024-2-9

Democratizing expertise: does the problem of false information change the calculus?*

Cathrine Holst**

Abstract

This increasingly salient risk of false information has bearings on philosophy too. The focus of the article is on the ongoing philosophical exchange on the idea of 'democratized' or of 'democratizing' expertise. The article starts out with presenting three philosophically grounded proposals regarding the democracy-expertise relationship: *science in democracy* – an approach primarily concerned with safeguarding independent scientific institutions positioned within a larger democratic system; *direct democratization* – an approach that focuses on expert arrangements more broadly and the need for direct measures of democratization; and *partisan expertise* – an approach which questions the possibility of independent, politically neutral expertise. The next section of the article provides a preliminary assessment of these proposals, before sketching a preferable fourth option: *epistemically justified expertise* – an approach focused on facilitating democratization measures which can be defended on epistemic grounds. The final section elaborates on the problem of false information and discusses whether the preliminary proposal assessment stands, or should be revised, confronted with this problem. Contrary to what is frequently claimed, it is argued that the problem of false information does not alter the calculus significantly. The problem of false information raises complex causal questions, and urgent questions of policy and regulation. It may also raise philosophical questions, but the specific philosophical discussion of how to (re-)design democracy-expertise relations, seems largely untouched.

Summary: Introduction. – I. Democratization of expertise. – II. A preliminary assessment – and a preferable fourth option. – III. The problem of false information. – Works Cited.

* I am grateful for comments from an anonymous reviewer. The article has also profited from inputs from participants at the conferences *New Waves in the Philosophy of Epistemic Authority and Expert Testimony,* Technische Universität Dresden, 5-6 October 2023, and *The Democratic Containment of Fake News and Bad Beliefs*, Luiss Guido Carli University, Rome, 26-27 October 2023.

** ORCID: 0000-0002-2231-5826.

Introduction

It is often said that philosophy must respond to, or, at least, take account of, new societal trends, technological and political challenges. One such development or challenge is the problem of "false information", "emerging as the most severe global risk anticipated over the two next years", according to the World Economic Forum (WEF). [1]

This increasingly salient risk of false information may have bearings on a range of philosophical topics. The specific focus here will be on the ongoing philosophical exchange on the idea of 'democratized' or of 'democratizing' expertise, taking place within, and sometimes across, the philosophical subdisciplines of epistemology, political philosophy, and philosophy of science.

Present-day calls to make expert arrangements compatible with democratic requirements – or to *democratize expertise* – come largely from corners outside academia, including citizen initiatives and social movements, as well as governments, policymakers, and various knowledge-brokers eager to provide science and expert advice with 'impact' and 'legitimacy' (Krick and Holst 2024). However, also academic studies and discourse – and philosophy specifically – seek increasingly to take democratic demands on expert communities and expert bodies into account. Unsurprisingly, when philosophers do so, they do not necessarily agree on what such demands mean and imply. In line with this, a set of different philosophically informed institutional proposals of how to (re-)design the democracy-expertise relationship are circulating.

In what follows, [2] a first section will outline three such proposals (I). They are, in short, *science in democracy* – an approach primarily concerned with safeguarding independent scientific institutions positioned within a larger democratic system; *direct democratization* – an approach that focuses on expert arrangements more broadly and the need for direct measures of democratization; and *partisan expertise* – an approach which questions the possibility of independent, politically neutral expertise.

The proposals are to be understood as ideal types – that is, they are not copied from specific interventions, but condensed versions of possible ways of institutionalizing the democratization of expertise. Still, as references will il-

[1] See their *Global Risk Report 2024* The report was brought to my attention by Sjøflot (2024).

[2] The first sections of this paper rely on a more elaborate presentation in Holst (2025). Note that a fourth proposal – *citizens as knowers* – is left out from the current paper not to overburden discussions.

luminate, elements in each of them speak to contributions and considerable trends in contemporary philosophical literatures. The proposals will furthermore be presented with reference to a set of claims and assumptions which philosophers have different views on – referred to, accordingly, as *philosophical variables*. Specifically, the claims and assumptions in question concern (1) epistemic criteria; (2) expertise reliance; (3) democratic legitimacy; (4) ethical and political values; and (5) implementation. These variables will be outlined in more detail at the very beginning.

The section that follows will make a preliminary assessment of the three proposals and sketch a preferable fourth option: *epistemically justified expertise* – an approach focused on facilitating democratization measures which can be defended on epistemic grounds (III).

Then enters the problem of false information (IV). The problem will be briefly elaborated upon before a discussion of whether the preliminary proposal assessment stands, or should be revised, confronted with this problem. Contrary to what is frequently claimed, it is argued that the problem of false information does not alter the calculus significantly. The problem of false information raises complex causal questions, and urgent questions of policy and regulation. It may also raise philosophical questions, but the specific philosophical discussion of how to (re-)design democracy-expertise relations, seems largely untouched.

I. Democratization of expertise

Philosophical variables

(1) Generally, expert arrangements have tended to be scrutinized and evaluated based on various *epistemic criteria*. This is clearly the case in epistemology, where central discussions circle around which more detailed epistemic terms and parameters to rely on in accounts of proper expertise (see Grundmann 2024 for a recent overview). Within philosophy of science as well scientific norms of inquiry and cognitive values are typically placed center stage. Similarly, political philosopher Thomas Christiano (2012, 31, 41) connects expert arrangements specifically to "the epistemic function" of political systems, and describes expertise as a "filter" to ensure the "truth sensitivity" of political decisions.

This common emphasize on epistemic standards and credentials raise the general question of how institutional proposals of democratizing expertise re-

late to them. Under such proposals, are epistemic criteria still considered primary? Or is rather democratization given priority?

(2) Both epistemologists and philosophers of science have tended to regard 'domain' experts (Goldman 2011) or knowledgeable specialists, commonly with scientific or other professional training, as indispensable, and conceptualized contemporary societies as essentially *expertise reliant* and generally characterized by a cognitive division of labor (Kitcher 2011, 20), and significant epistemic asymmetries between those with and without domain expertise. Contributions in political philosophy also refer to this situation of epistemic interdependence and expertise reliance and inspired by John Rawls' idea of 'general facts' this reliance has been referred to as "the fact of expertise" [3] in modern societies (Holst and Molander 2017, 236).

This raises questions of the more detailed relationship between this 'fact' and its possible democratization. How far are public institutions dependent on advanced science and specialist knowledge? And to what extent can the epistemic asymmetries of contemporary knowledge societies be reduced or evened out, as a result of well-designed 'democratization' processes?

(3) Regarding *democratization* in the context of contemporary expert institutions, there seems to be certain agreed upon minimum requirements. First, at least to the extent that expert institutions are somehow related to the governmental apparatus, they must be democratic in the sense that their powers have been granted them through acts of democratic delegation. The people or their representatives, for instance in parliament, should decide the mandates and discretionary scope of e.g. expert agencies and other knowledge bureaucracies, as well as make legislation that regulate science advice mechanisms and academic institutions. When this is not in place, it seems unreasonable to talk about such arrangements as 'democratic' or 'democratized' even in a minimal sense. Second, a democratic commitment is related to the idea that citizens and their representatives are in "the driver's seat" of value considerations and choose "the basic aims that society is to pursue" (Christiano 2012, 33). Once more, a conception of 'democratic' or 'democratized' expertise must, as a minimum, recognize this fundamental role for the citizenry, and what it may imply for expert arrangements.

Still, other claims and assumptions about democratization and democracy could vary. One question is what democratization implies for expert institutions apart from their delegated powers being democratically authorized. What

[3] Rawls (1993, xviii) identifies a set of "general facts" that he considers to be characteristic of modern societies, the most basic of which is the fact of so-called "reasonable pluralism".

is needed in terms of e.g. representation, participation, transparency, and accountability? Another question is which procedures are proper when citizens make decisions regarding values and aims. A fundamental distinction is between democratic decision-making based on "aggregative" or "deliberative" procedures (Peter 2009, Ch. 2 and 3, see also Chambers 2023). Granted that citizens should be in 'the driver's seat' of 'basic' ethical and political judgments, could these judgments by tapped by aggregating citizens' de facto interests and preferences, as revealed by for instance election results or in polls, or should the primary concern of democrats rather be citizens' considered views after deliberating with others?

(4) The relationship between *value judgments* and expertise arrangements also touches upon other philosophical controversies. One thing is to say that non-expert citizens should have the democratic privilege to decide on fundamental moral concerns and political priorities; another to argue that expert inquiries can and should be kept free from such value considerations altogether. In philosophy of science the classical view was that scientific investigations should take place in ways that are 'value-free' in the sense that the justification of scientific theories should not depend on ethical or political considerations (see e.g. Haack 1998 on the 'context of justification'). Along similar lines, a not uncommon view in political theory and philosophy is that experts should provide facts and technical knowledge about 'means' whereas questions of 'ends' should be left for non-expert consideration and judgment (recently Fjørtoft 2024).

Others have contested both the possibility and desirability of such a strict separation (e.g. Douglas 2009, Lafont 2019, Ch. 1). Granted that citizens should be on top of ethical and political judgments, this raises the question of whether citizens or politicians should not be involved – in one way or the other – even in expert inquiries that seem largely technical (and in science's 'context of justification'). To the extent that ethical and political judgments are *at all* made within expert arrangements (whether in the 'context of justification' or in other contexts), this raises moreover additional questions. First, can value judgments be made more qualified and reasonable, and can there even be value expertise – some kind of "moral experts" (Singer 1972, 115) or "normative experts" (Lamb 2020, 910) – and if so, should such experts play a role in contemporary expert arrangements, within the confines of citizens' 'basic' aims-setting? Or do democratic concerns speak against it? Second, in discussions of values; inside and outside expert arrangements, is some reasonable consensus possible, typically on fundamental constitutional requirements and moral norms, or does ethical and political pluralism run deep (e.g. Lafont 2019, 34-62, see also Rawls 1993)?

(5) Finally, regarding *implementation,* interventions can assume a significant need for innovating and establishing novel arrangements. Alternatively, they could concentrate on how to revise and re-design legacy institutions, be it universities, the bureaucracy, or for instance political parties. Moreover, in the latter case, the focus can be on incremental revisions and improvements, or on more fundamental changes.

In short: Is it assumed that truly democratizing expertise require radical institutional innovation – or will the re-design of established arrangements do the job? And in such re-design, is the assumption one of modest revision or of more thorough transformation?

Philosophical proposals

First proposal: Science in democracy

Few denies that independence, in some sufficient amount, is needed for expert institutions to function properly (SAPEA 2019, British Academy 2024, Owens 2015, Oreskes 2019), but independence is not only one of more concerns but at the heart of the first institutional proposal to be considered – *science in democracy*: A general prescription under this proposal is for expert arrangements to be as independent as possible from political interference, economic interest, and promotion of social values. Also, this implies a superior role for institutional arrangements of knowledge provision at arm's length from government, such as autonomous research universities and independent science advice.

What triggers this consistent weight on independence and autonomy is the concern for epistemic quality (e.g. Dellsén 2020) – which independence is considered as conductive to, or a condition for – and, correspondingly, an emphasis on epistemic criteria generally. Furthermore, this proposal largely equates expertise with science (e.g. Glüer and Wikforss 2022) and emphasizes the essential role of academic and professional specialists within the broader cognitive division of labor in society – while also valuing the importance of science communication and for scientists with their privileged insights to enlighten the broader public.

In line with this, the notion of making expertise 'democratic' is thought of in relatively minimal terms. That is, as far as expert arrangements are public institutions, they should have delegated authority from democratic decisions, for instance in parliament, and the scope of this authority must be limited so

fundamental political priorities, value interpretation and ranking are made by the citizenry and people's representatives (i.e. Christiano 2012, see also Wikforss 2019).

However, under this first institutional proposal, more direct measures of democratization are limited. Importance can be granted for instance to ensuring transparency of argumentation, use of sources, and procedures in expert inquiries, but only as long as openness can be shown not to harm epistemic quality (on "benefits of secret deliberations", see Kogelmann 2021, 73). Similarly, there can be a concern for e.g. the significance of epistemic diversity in science, yet tensions between inter- and trans-disciplinarity and scientific excellence can be highlighted too (e.g. Jacobs 2014).

A fundamental contention moreover is that scientists and other experts should concentrate on factual and technical inquiries and advice. This may be connected to a defense of a doctrine of value-freedom (see Betz 2013), or of minimal interference from ethical and political considerations (e.g. Stamenkovic 2024). This implies to put aside issues that may trigger controversy over social values, or to remain neutral and impartial when such issues are addressed. The recommendation moreover is that scientists should avoid prescribing public policy (Lackey 2007, SAPEA 2019).

With this point of departure, questions of the validity of value assessments and prescriptions, and of whether there can be expertise in the moral domain, do not arise with any force. The same goes for discussions on whether a reasonable agreement on what is right or just can be reached among people with different ethical and political outlooks. The strategy confronted with value influences is more one of avoidance, than one of qualification and search for potential consensus on core moral or constitutional requirements.

However, sometimes the avoidance of value judgments – risk assessments, cost considerations, recommendations, etc. – is impossible, and needs to be addressed. When so, the approach is to identify "what the public actually values", as expressed in elections, and by parliamentarian majorities, or by means of "empirical investigations" (Schroeder 2022, 252), and to ensure that the values promoted by experts "align" with (Bennett 2022, Sections 4-5, Gundersen 2024), citizens' preferences and interests.

In questions of implementation, this first type of proposal will typically emphasize legacy institutions and so existing contemporary expert arrangements – from classical universities to science advice mechanisms and a knowledge-based civil service – and to improve on them so they function in accordance with what is conceived to be their normative purpose and function: maintaining and cultivating independence, excellence, and neutrality (e.g. Col-

lini 2013, Heath 2020). This may require substantive re-design of existing institutions in some settings (in cases of severely politicized bureaucracies; when expert bodies are embedded in and hampered by conflict of interests; universities are put under market pressures, etc.). In other settings, where expert institutions already are relatively well-functioning, the task is rather one of defending and protecting their integrity and the existing design.

Second proposal: direct democratization

Just as the first proposal, the second proposal – *direct democratization* – supports the idea of making expert institutions with a potential of becoming trusted and legitimate across value conflicts and the political spectrum. The importance of institutions' independence and integrity will also typically be highlighted, and various epistemic criteria, as well as the significance of scientific expertise and professional specialists. However, under this proposal varied types of non-scientific – e.g. local and experience-based – knowledge, is considered essential too, and disciplinary diversity and trans-disciplinarity among scientific experts is embraced (Collins and Evans 2007, Owens 2015).

Crucially, moreover, democratic demands on expert arrangements are met with various direct measures (Fisher 2009, Oreskes 2019). They could include measures to improve representation – be it descriptive representation based on for instance gender, culture, and geography (e.g. Irzik and Kurtulmus 2021), or substantive representation in terms of political views, or stakeholder interests (e.g. Intemann 2015). They could include various mechanisms to enhance lay participation, where ordinary citizens participate on par with other experts as 'lay experts' or 'citizen experts' or are consulted in deliberative fora or other lay assemblies, or at digital platforms (for examples, see Krick 2021). Democratic accountability is also regarded as vital – expert bodies must account for their inquiries, assessments, and prescriptions in relevant democratic fora, be it elected assemblies or broader publics (e.g. Landwehr and Wood 2019). Transparency too is a core concern and may take radical shapes, for instance in terms of opening all sequences of expert inquiries and deliberations to the public, including mass and social media, and taking active measures to make background documents available to all affected (see Elliott 2020).

Furthermore, whereas the first institutional proposal will emphasize the need for scientists to communicate their findings to a broader public, this proposal rather highlights the need for intermediary institutions – from public media to social movements and non-governmental organizations. Such organizations work actively at the interface between science, policy-making, and the

public sphere, and by means of various communicational strategies and participatory and educational devices, to reduce epistemic asymmetries and ensure reciprocal exchange across the lay-expert divide (moving beyond top-down dissemination of a "linear model" perceived to be outdated, e.g. Fischer 2009, Owens 2015, Krick 2021). Also, whereas the first proposal prescribes for experts' value ranking– in as far as such ranking cannot be avoided – to align with de facto public values and ethical and political priorities in the citizenry, this proposal subscribes to some variant of consensus-oriented deliberative democracy, that is, to the idea that value commitments can be changed in processes of discourse and learning, and that it is possible for citizens with diverse ethical and political views to reach some agreement on fundamental norms and rights in inclusive processes of deliberative will- and opinion formation (recently Lusk 2021, see also Kitcher 2011).

It is essential under this second proposal, that facts and values, means and ends, technical and moral considerations, etc., are regarded as intertwined, and that interpretations and assessments of social values will affect expert inquiries in all contexts. This basic contention is what largely drives the strong impetus to democratize expert communities and institutions, as citizens should be on top of important ethical and political decisions. In contrast to what is assumed under the first proposal, such decisions cannot by and large be separated from factual and technical deliberations, and left for other arenas, but will be made within expert arrangements too, and to a significant degree. Furthermore, an assumption of equal moral and political competence rules out the possibility of cultivating expertise on such considerations; we cannot talk about 'expertise' in the moral domain (e.g. Fischer 2009, Lamb 2020, Dowding 2024).

Lastly, this second proposal is open to both the innovation of novel democratized expert institutions and to reform and re-design of legacy arrangements yet envisions overall a landscape of expert arrangements significantly different from the current one. In the transformed terrain imagined, established knowledge bureaucracies, advisory mechanisms, universities, and research institutes operate side by side with new democratized mechanisms and fora, and all institutions and arenas are consistently organized in accordance with a democratic ethos and aim at cultivating interactions and dialogue on equal terms across the lay-expert divide.

Third proposal: Partisan expertise

In contrast to the two previous proposals, the third proposal questions the ambition to cultivate expert institutions with legitimacy and support across constituencies, and the possibility and desirability of making such institutions neutral and independent from political and social interests. Accordingly, the main strategy is that of designing and re-designing for well-functioning *partisan expertise* – "expertise developed within the parameters of certain political commitments" (White 2024, 271) – and where the aim thus is not necessarily to make expert arrangements 'impartial' or 'neutral' (see also Rolin 2021). Such arrangements can take the shape of advocacy think tanks or be found as expert advisory structures within political parties, or civil society organizations, but can also take the shape of knowledge bureaucracies which are not developed as depoliticized knowledge bodies, but where recruitment considers political views and ideological orientation along with other merits and competences. Similarly, there would be research universities and institutes, but inhabited by academics and scholars often active within political parties or as movement intellectuals and pursuing research projects with political-ideological ambitions.

This is not to say that various epistemic criteria, scientific knowledge, and recruitment based on academic merits, are disregarded, but under this proposal – and differing from the first proposal – other concerns and criteria play a central role too, as epistemic standards and social interests, cognitive and non-cognitive values, science and politics are conceived of as interwoven.

At the same time – differing from the second proposal – this proposal pays less attention to a range of more direct measures to democratize expert arrangements. That is, the primary emphasis is not on developing mechanisms to energize lay participation; on implementing e.g. quota devices to ensure women, minority, etc. representation; or on radical transparency measures to expose the backstage of expert scrutiny and deliberations for the wider public. Decisive however is accountability towards relevant political fora. At the outset, such fora must delegate and define the proper mandates for the expert arrangement in question – be it when a parliamentary majority lays down terms of reference for a governmental advisory mechanism or decide on the policy for public universities; a party congress mandates the knowledge and analysis unit at the party office; or a social movement makes guidelines for its scientific advisory council. Yet in addition these knowledge structures, institutes, units, mechanisms etc. must report back and be held to account by the political principal in question: Members or member representatives in political organi-

zations, movements and parties must be given real opportunities to check on priorities within ongoing knowledge production; parliamentarian majorities and their government must actively manage the production and provision of expertise in the civil service and public bureaucracies (e.g. Downey 2021); research universities may have relative independence and seek scientific excellence, but must also adapt to changing priorities with shifting majorities and governments, etc.

This basic contention – that expert inquiries will be embedded in political priorities and value assessments, and that discussions of the design of expert institutions must profoundly relate to this – also brings to the fore where this proposal and the second proposal have similar features and contrast fundamentally with the first proposal. Still, whereas the second proposal responds to this state of affairs by means of a range of direct measures to democratize and emphasizes the potential of consensus-making through deliberation and of bringing experts and non-experts into interaction on equal terms, this third proposal rather emphasizes how disagreements often run deep and mirror substantial ideological and political divides and social and moral conflicts (e.g. Bellamy 2007, 191). Hence, building majorities, including by producing and utilizing knowledge and expertise, become decisive, as consensus may not be in reach (see also Rolin 2021, Hilligardt 2023). Moreover, instead of seeking to shrink cognitive inequalities by means of democratization measures and various bridging strategies, emphasis is put on cultivating and advancing – always partisan and political – expertise; on recruiting the better qualified, yet on the assumption that these experts will never be completely 'impartial' or 'neutral'; and on sharply checking on them, demanding accounts, and having experts sanctioned when needed (e.g. block reappointment when performance is 'bad', and re-appoint in cases of good performance).

Finally, under this proposal, the focus is not primarily on making new devices, but on reform of legacy institutions broadly speaking, from research institutions, and public bureaucracies, to political parties (on the latter, see Ebeling and Wolkenstein 2017). Reform efforts must furthermore be considerable, in line with the overall approach of embracing and cultivating expertise as competent partisanship, replacing what is perceived to be the current grammar of 'de-politicizaton' with a grammar of politicization or re-politicization.

II. A preliminary assessment – and a preferable fourth option

We approach the question of assessing these three proposals, by providing first a *prima facie* assessment of the claims and assumptions (i.e. variables 1-5) that underlie them.

Starting up with the role of epistemic criteria (*variable 1*), a case can be made for making such criteria primary in assessments of expert arrangements. Four (in part interrelated) arguments speak in favor of this. First, at least a considerable (if not primary) role for such criteria is in the end assumed under all the outlined proposals (*common denominator argument*). The science in democracy-proposal regards the epistemic function of expert institutions as fundamental. The other two proposals emphasize non-epistemic parameters too, such as representativeness and participation (direct democratization-proposal) or political or ideological criteria (partisan expertise-proposal), yet also recognize a significant role for cognitive criteria and scientific merits.

The approach of these two proposals highlights thus the question of how to rank the various criteria in circumstances when they may come into conflict, for instance when including lay persons on equal footing in expert deliberation results in undue and disproportional consideration of arguments that are irrelevant or obviously invalid; when transparency triggers public and media exposure with a chilling effect on experts' inquiries; or when political-ideological alignment trumps relevant domain competence in expert recruitment (see Kogelmann 2021, Christensen, Holst and Molander 2022, Ch. 6). Plausibly, in these situations, where non-epistemic and epistemic criteria are in tension, the latter type should take priority. This is first and foremost to ensure the normative legitimacy of political rule (*normative argument)*. For government or governance to have such legitimacy (Scharpf 1999, see also Chambers 2023), they must fulfill procedural democratic standards in terms of citizens' equal opportunities for participation in political decision-making and collective value judgments, but arguably they should also have some instrumental value in that they contribute to generally preferable outcomes, which would include decisions that are epistemically sound: well-informed, well-founded, etc. (e.g. Peter 2009). This granted, political rule must be institutionalized in ways that ensure such epistemic qualities to occur too, and it is hard to see how this can happen without some expert institutions which consistently prioritize epistemic credentials before political-ideological and participatory concerns.

The fact that several such institutions in contemporary societies have a de facto emphasis on the primacy of epistemic criteria in their mandate and guidelines (e.g. Gundersen and Holst 2022, Gundersen 2024) could reflect a

common understanding of this condition of normative legitimacy, or at least indicates that it seems hard to justify non-epistemic functions as primary for such institutions for the public eye (*public justifiability argument*).

Drawing in the same direction (see e.g. Rothstein 2011, Yesilkagit et al 2024) are the relatively high levels of trust among citizens in expert arrangements – from universities to knowledge bureaucracies – that are perceived to be competent, impartial, and fair (*public trust argument*). To be sure, real-world expert institutions will tend to play other roles too (see e.g. Boswell 2008): Expert arrangements can be utilized for strategic political purposes, function as negotiating arenas for conflicting social interests, or be sites of influence for affected stakeholders and ordinary citizens. However, when these other non-epistemic functions take center stage, and epistemic credentials and criteria are put aside, it seems that it will be difficult to justify them publicly and have citizens trusting them qua expert arrangements.

Regarding the condition of expertise reliance (*variable 2*), we see once more that all proposals, although variably, recognize non-experts' dependence on specialist scientific and professional knowledge in many contexts, and a non-trivial role for such knowledge in contemporary public policy and society (*common denominator argument)*. In the case of the first and third proposals (science in democracy-proposal and partisan expertise proposal), this comes clearly and explicitly to the fore: Under these proposals, there are experts and non-experts (whether or not expert knowledge is considered 'political' or more 'neutral'); we need to rely on these domain experts, and scientists in particular, to have sound knowledge and make good decisions; and even if experts should seek to communicate their expertise (first proposal) and are to be held to account by political principals and constituencies (third proposal), knowledge landscapes in modern societies will inevitably be characterized by considerable epistemic asymmetries. However, also the second proposal (of direct democratization) recognizes a certain cognitive division of labor where scientists and other professionals have a considerable role, even if it is assumed that other types of knowledge can be of equal importance and validity, and cognitive inequalities can be largely overcome by means of bridging strategies and intermediary institutions.

This common minimum acceptance of contemporary societies' expertise dependence across proposals, is related to a broader consensus within the human and social sciences of the fact of this dependence, as well as our everyday experiences of reliance on experts and expertise (*general fact argument*). This makes 'the fact of expertise' (Holst and Molander 2017) in some version hard to get around when formulating sensible proposals of institutional design, and

if proposals of democratizing expertise underplay this fact, or seem to disregard implications of it, this would be a reason to doubt them.

With the scope of a sound approach to variable1 (and variable 2) more settled, the more reasonable approaches to the different sub-questions under *variables 3 and 4* can be identified as well. Arguably, as epistemic criteria are primary for the assessment and design of expert arrangements, a condition for introducing various democratization measures would be that they are likely to contribute to higher epistemic quality, or at least does not significantly increase the risk of poorer epistemic outcomes. Hence, measures of inclusion and participation that are known to have considerable epistemic costs, or that increase the likelihood of such costs significantly, should be avoided.

Many such measures could however be allowed for on these terms. Ensuring transparency of expert sources and arguments is likely to facilitate critical scrutiny and control, and so improve on inquiries. Several measures supporting competent pluralism in expert arrangements are also unlikely to cause epistemic harm, and may even increase epistemic credentials, given what studies have shown so far of a positive relationship between cognitive diversity among inquirers and investigatory and deliberative quality (e.g. Sunstein and Hastie 2015, Moore and MacKenzie 2020). The latter speaks generally in favor of a relevant multi-disciplinary composition of expert communities and expert bodies. However, it also speaks for the inclusion of experts with varied demographic characteristics, or which possess vital stakeholder and local knowledge. These are all examples of epistemic diversity which can contribute to increasing the pool of information and arguments and sharpen investigations. Installing accountability procedures that oblige experts to explain their priorities and conclusions – be it in fora of peers or broader assemblies – can facilitate scrutiny, control, and quality too.

An additional virtue of such measures – of transparency, representation, participation, and accountability – is furthermore that they not only have epistemic credentials (when designed with care), but also that they may be conductive to more equal opportunity structures in public institutions, and as such they are arguably good also for procedural democratic reasons.

Regarding 'value-freedom', if it is the case that ethical and political considerations, in contrast to for instance scientific theories and claims, are not considerations that can be improved on from the perspective of epistemic criteria (i.e. 1), this would initially speak in favor of seeking to avoid them. However, even if it may be the case that some sequences of expert inquiries and deliberations (e.g. 'the context of justification') can be held 'value-free' (an assumption many would contest), this is not decisive, as such inquiries and

deliberations within expert institutions cannot avoid the inclusion of ethical and political judgments altogether, for instance when selecting which questions and topics to investigate and scrutinize, interpreting findings, or anticipating risks and costs.

Accordingly, since expert institutions are involved in such judgments, the focus should be on these judgements' epistemic credentials – if indeed judgments of this kind are such that they can be deemed better or worse epistemically speaking. Many philosophers would agree that not only claims of what is, but also of what ought to be, are possible objects of some kind of rational discourse (Gesang 2010, see also Hoffman 2012). Normative views do not simply reflect subjective beliefs and desires beyond reasonableness considerations, but may be better or more poorly qualified, justified, etc. This granted (and putting aside many philosophical intricacies), there are at least three implications for proposals of how to (re-)design expert institutions. First, if some ethical and political judgments are better than others, some persons may be better at making them, in the sense that they have skills or 'expertise' that are conductive to argumentative quality in this area. These 'moral' or 'normative' experts could be thought of as knowledgeable specialists – qualified for instance by means of philosophical training – or their qualification could take place in other ways (e.g. Hegstad 2024 on various models of ethics commissions). Either way, it would be good for expert arrangements to consider including them in their ranks.

Still, in democracies many normative considerations would and should be delivered to expert communities and expert bodies from the citizenry. Hence, second, if we want epistemically well-functioning expert arrangements, there is a case for preferring proposals for (re-)design that take a deliberative approach to democratic will- and opinion-formation, assuming that citizens' preferences can be cultivated in processes of discourse and learning.

Third, as it can be assumed neither that the outcome of such democratic deliberations will be consensus, nor that disagreement on ethical and political issues go deep, proposals should build in that both outcomes are possible, and institutionalize strategies for expert arrangements to handle them.

Finally, when it comes to implementation (*variable 5*) a case can be made, rather uncontroversially, for considering modest re-design of existing institutions before more radical re-design or design of new institutions, as such paths of institutional change and reform are *prima facie* likely to be more cost effective and have higher public support. Yet obviously, this is so only when existing institutions are already relatively well-functioning, and if not, more radical change and novel institutions may be defensible or needed.

With this said regarding variables 1-5, the following can be derived about the three ideal type proposals of possible ways of democratizing expertise. The first *science in democracy*-proposal is initially promising in that it puts epistemic criteria first and fully recognizes 'the fact of expertise' and the asymmetries that follow. It also allows for some more direct measures of democratization if they clearly support an epistemic purpose and suggests an incremental reform path with the declared aim of building on best practice in existing expert arrangements. However, this proposal also underestimates the democratic importance of making expert institutions more representative, participatory, accountable, etc. when this is compatible with epistemic concerns, but also the extent to which many such measures may have adequate epistemic credentials, if designed consciously with this in mind. The proposal at the same time puts too much emphasis on the option of avoiding ethical and political value judgment in expertise production and provision, assuming mistakenly that contemporary science advice mechanisms in their best version achieves 'value-freedom'. It also disregards the possibility of making value judgments more well-founded and robust, by means of deliberative-democratic decision- and consensus-making and by including experts with competences conductive to well-founded normative discussion.

The second *direct democratization*-proposal has a flexible and pragmatic approach to questions of implementation and design and an ambitious and laudable agenda of democratization and deliberative consensus-seeking and of establishing intermediary institutions between experts, policy-making, and the public. Still, the proposal lacks a consistent focus on epistemic criteria, and underestimates the extent to which democratic and epistemic concerns may be in tension and how there may be asymmetries between experts and non-experts that cannot be easily bridged. It also fails to take properly into account the extent to which citizens may reasonably disagree and legitimately conflict over ethical and political issues. This proposal furthermore disregards how deliberation within expert arrangements on value interpretation and ranking can be made more qualified and justified, and how recruitment and cultivation of expertise must consider this.

The third *partisan expertise*-proposal recognizes the cognitive division of labor and our reliance on experts and expertise in contemporary societies and focuses importantly on the role of knowledge production and provision in legacy institutions such as political parties and governmental organizations. This proposal is also right in recognizing how expert inquiries and deliberations are considerably influenced by ethical and political values, and how citizens and politicians may conflict deeply along political and ideological dimensions.

Still, it is a significant limitation of this proposal that it insufficiently recognizes the crucial role of independent expert arrangements – such as universities and research institutes, science advice mechanisms, or knowledge bureaucracies – that seek to put epistemic criteria first and before political-ideological considerations. It also seems to rule out the possibility that expert institutions can seek to approach value considerations impartially and competently, and that democratic processes can come with deliberative features that facilitate learning and consensus-making across constituencies.

Based on these considerations, a fourth proposal arguably suggests itself – *epistemically justified expertise* –similar in some respect to one or more of the previous three, yet avoiding their shortcomings, and so overall preferable to all of them. This proposal puts consistently epistemic criteria first and accepts our reliance on expert arrangements as a 'general fact', and what this fact implies in terms of unavoidable cognitive inequalities between domain experts and non-experts. It also grants science a special epistemic role in contemporary knowledge societies, yet recognizes the importance of different types of expertise, complementing science.

As a core feature, this proposal embraces measures to increase representation, participation, transparency, and accountability in expert arrangements, but only to the extent that such measures are unlikely to be epistemically harmful or significantly increase epistemic costs. It supports strategies and organizations to bridge expertise, policy-making, and the public too, yet consciously avoids aims-setting and recommendations which assumes that advanced specialists – in say engineering, virology or macroeconomics – and lay people without similar domain expertise could seek to operate as cognitive peers.

This fifth proposal focuses furthermore on how ethical and political considerations that are de facto made in expert arrangements can be justified and made as well-founded as possible, and on ensuring, by means of recruitment and qualification, that expert communities and bodies include inquirers with skills and competence in enhancing investigations of the value-laden issues and questions at stake. In this sense this proposal speaks of – and recommends the inclusion of – relevant 'moral' or 'normative' expertise.

However, as ethical and political decisions are mainly to be made by citizens in democracies, the proposal firmly recognizes this, but emphasizes a deliberative approach to the making of such decisions by the citizenry, assuming that citizens' interest and preferences could be made sensitive to and transformed by collective processes of communication and argumentation. Such deliberative processes regarding value interpretation and ranking, whether tak-

ing place among experts or non-experts, will sometimes end in consensus, for instance regarding some fundamental moral concerns, or constitutional norms. Other times disagreements stick, and decisions must be made by means of some other mechanism (e.g. voting or negotiations), and institutional solutions, this fifth proposal contends, must take both possibilities into account.

Finally, the proposal takes incrementalism as default, focusing consistently on identifying relatively well-functioning expert institutions in the existing landscape, based on adequate analyses of their empirical characteristics and effects (rather than on assumption or myth), and on revising and reforming them modestly. Yet, if epistemic concerns speak in favor of it, more radical reform would be needed too. The proposal thus speaks in favor of the innovation and design of novel organizations, bodies. and panels in the face of institutional gaps and deficiencies in existing arrangements, and when reform of legacy institutions will not do. (III)

III. The problem of false information

The question to be addressed is whether the problem of false information alters the preliminary assessment made in section II, including the conclusion that the sketched fourth option is preferable to the three initially outlined proposals.

The problem has many names – is has been referred to for instance as the challenge of 'disinformation' (Oreskes 2019), 'knowledge resistance' (Wikforss 2019), or 'post-truth politics' (Christensen, Holst & Molander 2022). It has also spurred philosophical and other debates which zoom in on specific types of false information or beliefs with arguably growing prevalence in contemporary societies and politics. Levy (2021, xi) for instance concentrates on a sub-category of false beliefs that he refers to as "bad beliefs" which are beliefs conflicting with "the beliefs held by the relevant epistemic authorities" and "held despite the widespread public availability either of the evidence that supports more accurate beliefs or of the knowledge that the relevant authorities believe as they do". Examples are rejection of "climate change, in defiance of the scientific authorities", or of "vaccines, in defiance of the medical profession" Others have focused on the partly overlapping phenomenon of 'conspiracy theories' (e.g. Cassam 2019) or on 'fake news' (Zimdars & McLeod 2020).

Across diverging approaches and treatments, most agree however on the severe, and possibly devastating consequences of the problem of false infor-

mation. Illustratively, the WEF report elaborates on how "misinformation and disinformation may radically disrupt electoral processes"; and how "a growing distrust of information", and of "media and governments", may "deepen polarized views" and spur "a vicious cycle that could trigger civil unrest". Importantly, we might also see "a risk of repression and erosion of rights as authorities seek to crack down on the proliferation of false information" (see paragraph 1.3 in the report).

Faced with this super-challenge and 'most severe global risk', as noted by the WEF, also philosophers have been engaged, and returning to our topic – the democracy and expertise relationship – strong claims are made – about the virtues (or flaws) of this or the other proposals when it comes to dealing with the false information challenge specifically– that at a first glance might imply a re-assessment of the philosophically informed institutional proposals. If so, and given the graveness of this challenge, our preliminary considerations in the previous section might not hold.

There is, first, a way of approaching the problem of false information that proponents claim strengthens the case for something along the lines of the science in democracy proposal (see e.g. Wikforss 2019, Glüer an Wikforss 2022). The problem thus conceived emphasizes how false information occurs and spreads; on the one hand, because the independence and integrity of expert communities and institutions, and science especially, are disrespected or under pressure, endangering the epistemic quality of the presumably most authoritative knowledge production, and increasing the chance of falsehoods occurring or not being corrected; on the other hand, because the up-take of validated scientific knowledge and the scientific consensus in the broader society, including in segments where false and 'bad' beliefs are likely to thrive, is blocked, disturbed, or not ensured properly. Accordingly, measures to be taken, according to this conception of the problem of false information, focus in part on how to protect and develop sufficiently autonomous scientific institutions and expert communities, where recruitment take place according to competence and merit, and inquiries and deliberations are not affected unduly by political and media pressures or economic and commercial interests. In part, measures are proposed that would target how science and generally the best available knowledge spreads from expert environments to non-experts, to ensure proper up-take. Levy (2021) elaborates here on two paths. The standard approach would be science communication broadly speaking, where scientists disseminate their consensus findings through a well-functioning public sphere, including aptly regulated social media platforms and interactive technologies, and assisted by an education system where science education and critical

thinking are put central stage (see also e.g. Cassam 2019). The other path, preferred by Levy, is to add harder measures – and a regime of 'epistemic engineering' based on 'nudging' – that better recognize, as he sees it, the level of 'pollution' and 'manipulation' characteristic of contemporary epistemic environments.

However, even if this approach and (somewhat varied) conception(s) of the false information problem could speak in favor of something like the science in democracy-design, the sketched fourth option could arguably do the job too. Also under this proposal, both the independence and integrity of science and expert institutions needed to ensure epistemic quality, and proper uptake among citizens and non-experts of the beliefs validated by such institutions are vital, and measures proven to ensure it would be supported.

There are similarly – and secondly – approaches to the problem of false information that supporters say speak in favor of (re-)design proposals along the lines of the direct democratization ideal type (e.g. EGE 2023, 2024, Oreskes 2019, Krick 2021, Geissel 2024). The problem thus conceived emphasizes how facts and values are intertwined, and so how inquires in expert institutions and science to weed out false theories and beliefs always will involve ethical and political considerations too. Hence, since non-epistemic value judgments are generally up to citizens in a democracy and not something we delegate to scientists, and citizens are unlikely to trust outcomes of processes from which they are unduly excluded, this speaks for a range of direct measures of expertise democratization and for the making of various intermediary institutions of knowledge brokering and science-society dialogue, and for generally connecting the problem of false information, and how to address it, to broader efforts of democratizing public institutions.

Yet, once more, also under this conception of the problem of false information, the fourth alternative would seem to work equally well. This alternative also recommends measures of democratizing expert bodies – although in ways that are compatible with the epistemic rationale of such bodies – and supports broader democratization efforts at the science-policy interface and in public life to ensure political equality and citizens' trust.

Third, some have claimed too that the problem of false information is better addressed by an approach that shares traits with the partisan expertise ideal type (e.g. Ebeling and Wolkenstein 2017, White 2024). Once more, we are presented for a conception of the false information problem which emphasizes how assessments of 'information' inevitably will involve value considerations, whether such assessments take place inside or outside expert institutions. And once more, proponents highlight how this fact gives us reasons to 'democra-

tize', not in the sense of direct measures democratizing expert bodies and scientific communities, but in the sense of building up, not impartial, but explicitly partisan expertise structures within political parties, social movements and civil society, that properly take into account and are intertwined with the political value conflicts and interest struggles of our time. Finally, and again, there are democratic reasons for this move, according to proponents, but the approach is also key to ensure citizens' trust in public institutions and knowledge-building and should hence be at the core of any strategy to address falsities, fake news and conspiracy theories.

However, yet again, even this description of the problem of false information could be adapted by the fourth suggested ideal type. Even this alternative would support and encourage the institutionalization of hybrid structures of knowledge and expertise in politics and civil society, even if it insists at the same time for the need and primacy epistemically speaking of independent and impartial structures of science and expertise and would highlight how well-functioning partisan expertise in the end are reliant on trustworthy input from such more impartial structures.

Hence, as it turns out, the preliminary assessment of our presented ideal types of the democracy-expertise relationship – and the case for the sketched fourth approach – in fact seems to hold when scrutinized from the perspective of the false information problem, and the calculus need thus no revision, at least not for this specific reason.

This is not to deny the severe societal effects of the problem at hand, and how it has challenged profoundly, both scholarship directed towards causal analysis (how to explain the rise – and the speed of the rise – of this problem?), and regulatory discussions (how to address the problem in policy and regulation?). Certainly, there is also no intention of denying that the problem gives rise to new important philosophical discussions, including within epistemology, but as for the exact exchange scrutinized in this article – of how (and how not) expertise should be democratized –assessments would rather seem to hinge on other issues.

Works Cited

Anderson, E. 1995. "The Democratic University: The Role of Justice in the Production of Knowledge." *Social Philosophy & Policy* 12 (2): 186-219.

Bennett, M. 2022. "Judging Expert Trustworthiness: The Difference Between Believing and Following the Science." *Social Epistemology* 36 (5): 550-560.

Betz, G. 2013. “In Defence of the Value-Free Ideal.” *European Journal for Philosophy of Science* 3: 207-220.
Boswell, C. 2008. “The Political Functions of Expert Knowledge: Knowledge and Legitimation in European Union Immigration Policy.” *Journal of European Public Policy* 15 (4): 471-488.
British Academy. 2024. *Public Trust in Science-for-Policymaking.* Report to the Prime Minister’s Council for Science and Technology.
Chambers, S. 2023. *Contemporary Democratic Theory.* Polity.
Christensen, J., C. Holst, and A. Molander. 2022. *Expertise, Policy-Making and Democracy.* Routledge.
Christiano, T. 2012. “Rational Deliberation Among Experts and Citizens.” In *Deliberative Systems: Deliberative Democracy at the Large Scale*, edited by J. Parkinson and J. Mansbridge. Cambridge University Press.
Collini, S. 2013. *What Are Universities For?* Penguin.
Collins, H., and R. Evans. 2007. *Rethinking Expertise.* University of Chicago Press.
Dellsén, F. 2020. “The Epistemic Value of Expert Autonomy.” *Philosophy and Phenomenological Research* 100: 344-361.
Douglas, H. 2009. *Science, Policy, and the Value-Free Ideal.* University of Pittsburgh Press.
Douglas, H. 2021. “The Role of Scientific Expertise in Democracy.” In *The Routledge Handbook of Political Epistemology*, edited by H. Douglas.
Dowding, K. 2024. “Ethical Expertise and Moral Authority.” *Res Publica* 30 (1): 31-46.
Downey, L. 2021. “Delegation in Democracy.” *Journal of Political Philosophy* 29 (3): 305-329.
Ebeling, M., and F. Wolkenstein. 2017. “Exercising Deliberative Agency in Deliberative Systems.” *Political Studies*: 635-650.
Elliott, K. C. 2020. “A Taxonomy of Transparency in Science.” *Canadian Journal of Philosophy*, 1-14.
European Group of Ethics in Science and New Technologies (EGE). 2023. *Democracy in a Digital Age.*
European Group of Ethics in Science and New Technologies (EGE). 2024. *Defending Democracy in Europe.*
Fischer, F. 2009. *Democracy & Expertise: Reorienting Policy Analysis.* Oxford University Press.
Fjørtoft, T. 2024. “Inductive Risk and the Legitimacy of Non-Majoritarian Institutions.” *British Journal of Political Science* 54 (2): 389-404.
Geissel, B. 2024. *The Future of Self-Governing, Thriving Democracies: Democratic Innovations By, With and For the People.* Routledge.
Gesang, B. 2010. “Are Moral Philosophers Moral Experts?” *Bioethics* 24 (4): 153-159.

Glüer, K., and Å. Wikforss. 2022. "What is Knowledge Resistance?" In *Knowledge Resistance in High Choice Information Environments*, edited by Jesper Strömbäck et al. Routledge.
Goldman, A. 2011. "Experts: Which Ones Should You Trust?" In *Social Epistemology: Essential Readings*, edited by A. Goldman and D. Whitcomb. Oxford University Press.
Grundmann, T. 2024. "Experts: Who Are They and How Can Lay People Identify Them?" In *Oxford Handbook of Social Epistemology*, edited by J. Lackey and A. McGlynn. Oxford University Press.
Gundersen, T. 2024. "Trustworthy Science Advice: The Case of Policy Recommendations." *Res Publica.*
Haack, S. 1998. *Manifesto of a Passionate Moderate: Unfashionable Essays.* The University of Chicago Press.
Heath, J. 2020. *The Machinery of Government.* Oxford University Press.
Hegstad, E. 2024. "Why Ethics Commissions? Four Normative Models." *Res Publica* 30: 67-85.
Herzog, L. 2023. *Citizen Knowledge. Markets, Experts, and the Infrastructure of Democracy.* Oxford University Press.
Hilligardt, H. 2023. "Partisan Science and the Democratic Legitimacy Ideal." *Synthese* 202: 135 (Online First).
Hoffman, M. 2012. "How to Identify Moral Experts? An Application of Goldman's Criteria for Expert Identification to the Domain of Morality." *Analyse & Kritik: Zeitschrift für Sozialtheorie* 34 (2): 299-313.
Holst, C., and A. Molander. 2017. "Public Deliberation and the Fact of Expertise: Making Experts Accountable." *Social Epistemology* 31 (3): 235-250.
Holst, C. 2025. "Democratizing Expertise: The Epistemic Track." *Social Epistemology* (forthcoming).
Intemann, K. 2015. "Distinguishing Between Legitimate and Illegitimate Values in Climate Modeling." *European Journal of Philosophy of Science* 5: 217-232.
Irzik, G., and F. Kurtulmus. 2019. "What Is Epistemic Public Trust in Science?" *The British Journal for the Philosophy of Science.*
Irzik, G., and F. Kurtulmus. 2021. "Well-Ordered Science and Public Trust in Science." *Synthese* 198: 4731-4748.
Jacobs, J. A. 2014. *In Defense of Disciplines: Interdisciplinarity and Specialization in the Research University.* The University of Chicago Press.
Kitcher, P. 2011. *Science in a Democratic Society.* Prometheus Books.
Kogelmann, B. 2021. "Secrecy and Transparency in Political Philosophy." *Philosophy Compass* 16 (4).
Krick, E. 2021. *Expertise and Participation.* London: Palgrave.
Krick, E., and C. Holst. 2024. "Democratising Expertise." In *Handbook of Policy Advice*, edited by Edgar Eldar.

Lackey, R. T. 2007. "Science, Scientists, and Policy Advocacy." *U.S. Environmental Protection Agency Papers*, Paper 142.
Lafont, C. 2019. *Democracy Without Shortcuts: A Participatory Conception of Deliberative Democracy.* Oxford University Press.
Lamb, R. 2020. "Political Philosophy and the Nature of Expertise." *Critical Review of International Social and Political Philosophy* 23 (7): 910-930.
Landwehr, C., and M. Wood. 2019. "Reconciling Credibility and Accountability." *European Politics and Society* 20 (1): 66-82.
Levy, N. 2021. *Bad Beliefs: Why They Happen to Good People.* Oxford University Press.
Lusk, G. 2021. "Does Democracy Require Value-Neutral Science? Analyzing the Legitimacy of Scientific Information in the Political Sphere." *Studies in the History and Philosophy of Science, Part 1* 90: 102-110.
Mansbridge, J., and S. Macedo. 2019. "Populism and Democratic Theory." *Annual Review of Law and Social Science* 15: 59-77.
Moore, A. 2017. *Critical Elitism: Deliberation, Democracy, and the Problem of Expertise.* Cambridge University Press.
Moore, A., and M. MacKenzie. 2020. "Policymaking During Crises: How Diversity and Disagreement Can Help Manage The Politics of Expert Advice." *The BMJ.*
Oreskes, N. 2019. *Why Trust Science?* Princeton University Press.
Pamuk, Z. 2021. *Politics and Expertise: How to Use Science in a Democratic Society.* Princeton University Press.
Owens, S. 2015. *Knowledge, Policy, and Expertise: The UK Royal Commission on Environmental Pollution 1970-2011.* Oxford University Press.
Peter, F. 2009. *Democratic Legitimacy.* Routledge.
Rawls, J. 1993. *Political Liberalism.* Columbia University Press.
Rolin, K. 2021. "Objectivity, Trust, and Social Responsibility." *Synthese* 199 (1-2): 513-533.
Rothstein, B. 2011. *The Quality of Government: Corruption, Social Trust, and Inequality in International Perspective.* The University of Chicago Press.
SAPEA. 2019. *Making Sense of Science for Policy Under Conditions of Complexity and Uncertainty.* Evidence Review Report No. 6.
Scharpf, F. 1999. *Regieren in Europa: Effektiv und demokratisch?* Campus Verlag.
Schroeder, S. A. 2022. "Thinking About Values in Science: Political Versus Ethical Approaches." *Canadian Journal of Philosophy* 52 (3): 246-255.
Shapiro, I. 2001. *Democratic Justice.* Yale University Press.
Singer, P. 1972. "Moral Experts." *Analysis* 32 (4): 115-117.
Sjøflot, O. 2024. "Reconsidering Echo Chambers and Epistemic Bubbles." Master thesis in philosophy. University of Oslo.
Stamenkovic, P. 2024. "Straightening the 'Value-Laden Turn': Minimising the Influence of Extra-Scientific Values in Science." *Synthese* 20.

Sunstein, C. R., and R. Hastie. 2015. *Wiser: Getting Beyond Groupthink to Make Groups Smarter.* Harvard Business Review Press.
White, J. 2024. "Technocratic Myopia: On the Pitfalls of Depoliticizing the Future." *European Journal of Social Theory* 27 (2).
Wikforss, Å. 2019. *Why Democracy: On Knowledge and Rule of the Many.* Fri Tanke.
Yesilkagit, K., et al. 2024. "The Guardian State: Strengthening the Public Service Against Democratic Backsliding." *Public Administration Review.*

DOI 10.17473/2240-7987-2024-2-10

The Collective Underpinnings of Bad Beliefs *

Säde Hormio**

Abstract

Even with events like the Capitol attack, it is misguided to focus too much on the possible epistemic failures of individuals. Instead, the focus should be on the collective underpinnings of bad beliefs (such as false beliefs about a stolen election), and especially on the collective agents who peddle in misinformation. We can divide the collective agents that pollute our epistemic neighborhoods roughly into those that do so for ideological or other such reasons (*misbelievers*), and those that do so for instrumental reasons (*disinformers*), although in practice these categories can overlap. These two motivations impact the responsibility of the collective agents that help to create bad epistemic neighborhoods. While misbelieving is more culpable in a purely epistemic sense, being a disinformer is more culpable in a moral sense. Epistemic institutions present a special case for the responsibility of collective agents. Although collective agents can present us with clear cases of culpability in epistemic matters, when dealing with consumers of fake news and misinformation, we should proceed with a certain level of epistemic humility.

* The work was supported by the Research Council of Finland under grant number 355849 and the Finnish Cultural Foundation under grant number 00230424. I would like to thank the anonymous reviewer for their helpful suggestions and comments. I would also like to thank the audiences at the 2023 workshops *The Democratic Containment of Fake News and Bad Beliefs* in Rome, the *Institutional Epistemology* in Helsinki and *The Role of Knowledge for Responsible Decision Making in Organizations* in Hamburg, as well as at the *Social Ontology 2024* conference at the Duke University. An early (and quite different) version of the paper benefited from questions from the audience members at the *American Philosophical Association (APA) Pacific Division Meeting 2021* and the *PoS Seminar* at the University of Helsinki, with special thanks for the written comments from Uskali Mäki and Samuli Reijula. I am also grateful to Don Fallis, Avram Hiller, Maria Lasonen-Aarnio and Holly M. Smith for their suggestions during a virtual meeting in June 2020, conducted to replace the original session at the cancelled *APA Pacific 2020*.

** ORCID: 0000-0002-2231-5826.

Summary: Introduction. – I. Bad epistemic neighborhoods. – II. Misbelievers and disinformers. – III. Responsibility of collective agents and consumers of fake news – Conclusion. – Works Cited.

Introduction

On 6 January 2021, thousands of Trump supporters gathered in Washington, D.C. for a rally in which the president encouraged people to march to the Capitol building, where members of the United States Congress were in the process of certifying election results. The rhetoric used by Trump framed the rally as a necessary response to a perceived stolen election, a narrative that he and his allies had repeatedly promulgated since he lost the election back in November 2020. The claims about widespread voter fraud and irregularities in the 2020 United States presidential election were not based on any evidence. According to a recent lawsuit, the numbers used to back up such claims were instead completely made up, with statistics about voter fraud fabricated by simply making up numbers out of thin air. [1] Although claims about widespread voter fraud were repeatedly refuted by election officials and independent observers, the false narrative about a stolen election gained traction among some segments of the population, contributing to false beliefs about the election being illegitimate. Following the rally, a mob of Trump supporters proceeded to the Capitol, where they had violent confrontations with law enforcement and caused significant property damage. The situation escalated rapidly, resulting in the invasion of the Capitol. The police force was overwhelmed, and the Capitol was placed under lockdown. The attack was widely perceived as an unprecedented assault on democratic processes, and it raised concerns about extremism in the United States, laying bare the need to protect democratic institutions. It also highlighted the potent force of misinformation. Some individuals lost their lives because of the attack. Through the propagation of false narratives about the election, and the amplification of these narratives by the right-wing media and social media, false beliefs contributed to real harms.

Beliefs are all around us. We hold innumerous beliefs about ourselves, each other, and how the world is. These beliefs come with various epistemic credentials. Bad beliefs are beliefs that are false, unjustified, or based on faulty reasoning. False beliefs do not correspond to facts or reality. Unjustified

[1] UNITED STATES OF AMERICA v. DONALD J. TRUMP, CRIMINAL NO. 23-cr-257 (TSC).

beliefs lack credence, that is, sufficient evidence or justification to be considered credible. Beliefs can also be irrational, based on faulty reasoning, and suffer from various biases. The kind of bad beliefs that I will focus on in this paper are beliefs based on misinformation and/or overtly dogmatic thinking. Misinformation usually refers to false information, but it can also be misleadingly and selectively shared accurate information (O'Connor & Weatherall 2019). When we act on beliefs with bad epistemic credentials, or make judgements based on them, we might behave in ways that cause harm or are contrary to how we would behave if we did not hold those beliefs. In addition, what we doubt or do not believe, or suspend judgement on, influences our decisions and actions.

It has been suggested that to counteract bad beliefs, we need to engineer epistemic environments to scaffold better beliefs (Levy 2022). [2] This paper agrees with the assessment and argues that epistemic environments are a matter for shared and collective responsibility. My focus is on the collective agents that can influence epistemic communities and make them epistemically bad in relation to some issue. Collective agents are organized collectives, such as corporations and governments. Their structure lays out their internal policies, procedures, and rules, and influences their collective character, all things which make it sensible to attribute responsibility to them.

Political discourses that demonstrate a disregard for facts and expertise pose undeniable epistemic problems for democracies (Rietdijk 2024). If we are worried about false beliefs for any kind of practical reasons – for example how they contribute to political polarization, or lack of support for meaningful policy measures to tackle important issues – then we must look at the collective level, as it will be the most relevant level in terms of responsibility. Consider someone who has received basic education only (poorly planned and executed at that) with the curriculum devoid of anything on media literacy. They live in a community where the 2020 United States presidential election result is considered a hoax, and this is also what their family and friends believe. Furthermore, they are surrounded by lies and misinformation around the issue, readily available to them via their favorite media sources and suggested to them by algorithms on the internet. Even if there is some degree of blame for the individual, the blame is better directed at the sources creating and contributing to the epistemic environment, such as media organizations and lobby groups.

[2] Levy's suggestion is not to manipulate anyone, but to rather give people reasons to change their minds by changing their epistemic environment.

I suggest that in cases where the epistemic environment plays an important role, the culpability for bad beliefs applies primarily not to the individual party holding such beliefs, but to the agents that cause or reinforce such bad beliefs or ignorance in direct or indirect ways. Examples include state propaganda during wartime, or corporations lobbying against regulation by spreading misleading claims about things like the harmfulness of tobacco. The election fraud narrative also falls into this stream. While questions around individual culpable ignorance or bad beliefs offer interesting epistemological puzzles in themselves, when it comes to issues like wide-spread denialism of some facts, it would not make sense to try to answer such questions in isolation from the epistemic neighborhoods. In such cases, the culpability for individual ignorance or bad beliefs is always a contextual matter.

The main contribution of this paper is normative: a discussion on how collective agents can be understood as culpable for individuals' bad beliefs or ignorance within an epistemic community, and more specifically, how the motivations of collective agents might affect their own culpability. I will explain why the motivations of collective agents are not private in the same sense as that of people and are thus relevant for evaluating the actions of collective agents in the public sphere. While I do not discuss potential real-life implications that identifying such responsibility should have in a legal sense or through other regulatory measures (although I do mention some such possibilities), judgements about the culpability of collective agents would clearly need to feature in any such account. [3] After all, there is nothing to replace the role of normative arguments when it comes to trying to answer the fundamental questions about what we should do and why, what is just and unjust, and how we should organise our lives together, including our epistemic lives. Normative arguments do not merely express personal preferences but, in the words of legal scholar Joseph W. Singer (2009), they are also "evaluative assertions and moral demands we are entitled to make of each other". Engaging with normative claims is thus about more than just stating one's opinions. They are imperative to any real debate about values. So, although I do not present ready-made solutions, clarifying the normative landscape is important, because at the end of the day, legal measures and policy measures both incorporate normative elements.

I begin by discussing the literature on the impact of epistemic environments on individual culpability. This will help to explain why the culpability

[3] For one suggestion on sanctioning epistemic agents that hinder an epistemic environment, see Ryan (2018).

of individual bad beliefs or ignorance is always a contextual matter. I then discuss the responsibility of collective agents, distinguishing between two different motivations for misinformation: mostly ideological or cynical. Both forms can have bad epistemic consequences and have contributed to the current political polarization. However, I will argue that while having bad beliefs based on ideological or other such reasons can be more culpable in a purely epistemic sense, the cynical form of denialism is morally more culpable. I finish by looking at the responsibilities of epistemic institutions and consumers of fake news.

I. Bad epistemic neighborhoods

Ignorance can be an excusing condition when assigning blame to an agent: had they known some relevant fact, they would have acted differently under the circumstances. Yet ignorance does not automatically excuse, and the subject can deserve blame for some ignorance and its consequences. An expert should have read the relevant literature pertaining to their area of specialization, but chose to do something else instead, resulting in a subsequent choice that was wrong or that led to bad consequences (Smith 1983). Culpable ignorance can just as easily relate to everyday life, like being late for an appointment due to a failure to check the address properly in advance. Absence of knowledge or lack of true belief can also be culpable sometimes, but I will not go into the debates over which conditions and under what circumstances this happens. What matters for my purposes is that if enough people are ignorant of important matters or hold false beliefs about them, it can lead to bad consequences. For example, research suggests that exposure to misinformation about climate change makes individuals less likely to want to take action to reduce emissions (Jolley and Douglas 2014), while exposure to conspiracy theories and misinformation about the election was arguably the reason why many took part in the Capitol attack. In short, bad beliefs can lead to bad consequences.

The question arises: are we responsible for trying to ensure that we do not hold bad beliefs? It cannot be the case that we have a responsibility to question each of our beliefs, even when we know that some of them must be untrue. From a psychological standpoint such a position would be untenable: we simply cannot go around questioning everything. Yet it also seems straightforward to say that sometimes we can be held responsible for holding bad beliefs. William Kingdon Clifford famously argued nearly 150 years ago that we

can be blameworthy if we form a belief based on insufficient evidence: if we have come to acquire such a belief because of stifling our doubts around the issue, failing to look into the issue properly, or by ignoring available evidence, we have no right to believe it (Nottelmann and Fessenbecker 2019). [4] Still, it would be unfeasible to demand that we question all our beliefs. William James (1896) noted as much in his response to Clifford, arguing that such a high standard in sufficient evidence would lead to paralysis in action, and that in contrast beliefs can be sufficiently justified based on personal or pragmatic reasons. He agrees with Clifford that one should first engage in inquiry, but if a decision should be made but cannot be made on evidence alone on a non-trivial matter, then one can decide on other grounds.

These days epistemological debates have mostly moved away from a focus on gathering first-order evidence towards issues such as whose testimony we should trust and how to identify expertise. Many things we believe in are based on what others have told us, including what science tells us. It is epistemically reasonable to believe that the trees I see outside my window exist in the courtyard of the building I work in, but it is also epistemically reasonable to believe that I live on the third planet from our universe's sun, although it is not a belief that I can formulate based on looking out of the window. Our epistemic circumstances affect what is epistemically reasonable for us to believe. Bad beliefs should be understood in the context of our current state of (scientific) knowledge about something. We need to defer to other people with expertise in the area in order to have good beliefs. Even so, it seems that stifling our doubts around an issue or failing to look into it properly qualifies for culpable ignorance or bad beliefs, at least if the matter has normative significance.

Apart from failing to investigate properly, agents can also be culpable for showing insufficient concern. Importantly, bad beliefs do not need to boil down to an individual's own epistemic failures only, as a whole community can hold false beliefs about an issue. There is an interesting body of literature on the effect that the environment where a person is raised in has on the culpability of their ignorance and on their responsibility (e.g. Arpaly 2002; Mason 2015; Wolf 1987). We could argue that if a community has failed to know something, this makes those who grow up in it less culpable in terms of their ignorance of that issue. In Nomy Arpaly's (2002, pp. 103-104) example, Sol-

[4] It is unclear if Clifford is making a moral or epistemic argument, or a combination of the two. Haack (1997) lists five different ways in which the relationship between ethics and epistemics could be understood in Clifford's argument.

omon grows up in an environment where women are thought to be less capable than men intellectually, especially when it comes to abstract thinking. Solomon's sexist belief is not irrational, as all the women in his community conform to the stereotype and no one questions it. Solomon leaves this small and isolated community when he receives a scholarship to an excellent academic institution. This means that he also leaves the epistemic environment in which only men are deemed capable of abstract thinking. Once outside his childhood community, Solomon meets outstanding female students and brilliant female professors. According to Arpaly (2002), he fails to respond to relevant moral concerns if he does not change his beliefs about women when confronted with counterevidence daily. By not changing his beliefs, he is irrational and displays bad will.

I give the label *bad epistemic neighborhoods* for those epistemic communities and environments in which you are likely to obtain false beliefs in relation to some issue. For example, if you live in a community where vaccine misinformation abounds, you live in a bad epistemic neighborhood in relation to vaccines. In a similar way, if a social media algorithm regularly suggests fake news about a political party to you, you are in a bad epistemic neighborhood in relation to that political party. Living in a bad epistemic neighborhood in relation to some issue *X* does not mean that all your beliefs are false. It only means that you are likely to hold false beliefs in relation to *X*, like Solomon did regarding the intellectual capacities of women.

Could there be an obligation for people to reach outside their epistemic communities in order to challenge their beliefs on normatively important issues? While we cannot directly control what we believe, we can take many actions that will indirectly influence our beliefs, like choosing to expose ourselves to new epistemic sources. Maybe epistemic agents could be blamed for not testing their beliefs, at least to some degree and when the stakes are high. While this sounds reasonable, there needs to be an outside spark of some kind to facilitate such questioning. After all, there are many beliefs that we do not even consider questioning in our daily lives. If you think that there is no need to examine an issue more carefully, it could be that agents are only partially responsible for their bad beliefs (Robichaud 2017).

You can easily imagine many kinds of epistemic neighborhoods in which questioning some belief, even with important normative force, would not even cross a person's mind. However, for a bad epistemic neighborhood to shield you from responsibility in relation to your bad beliefs about an issue, you must not have any inclination that it might be a bad epistemic neighborhood in relation to that issue. If you have a nagging worry that it might be and thus have

an incentive to not to investigate further to not to risk bursting your bubble, then your behavior could be blamed for a number of epistemic reasons (e.g. Piovarchy 2021). With any contentious or high-stakes issue, there is likely a spectrum of epistemic neighborhoods where in some you should question your beliefs, while in others the epistemic effect of the environment is just too strong to make demands on individuals. If an individual lives in an epistemically bad neighborhood in relation to some issue their whole lives (and the community does not change its epistemic outlook on the issue), we might disapprove of their beliefs, but we do not condemn them. Even so, we might want to condemn the neighborhood and those responsible for creating and upholding it.

Bad epistemic neighborhoods come in different varieties, depending on the issue. They can be real geographical areas or online communities, large or small, influential to outsiders or not, more or less robust, on so on. They can also take the form of media bubbles, like hyper-partisan news sources. Their grip on their members will also vary. In an influential paper, C. Thi Nguyen (2020) distinguishes between epistemic bubbles, in which certain relevant voices are omitted, possibly accidentally, and echo chambers, in which such voices are actively discredited. Solomon's childhood community could be described as an epistemic bubble, where some pertinent viewpoints have been excluded. These bubbles do not let in certain kinds of evidence, leading to "a social epistemic structure which has inadequate coverage through a process of exclusion by omission" (Nguyen 2020, p. 143).

In contrast, in an echo chamber, other voices are actively discredited, and the bad epistemic neighborhood is thus purposefully created and maintained. Echo chambers seek to epistemically discredit non-members, labelling them as unreliable or dishonest, while amplifying the epistemic credentials of members (along with a core set of beliefs), duly creating "a significant disparity in trust between members and non-members" (Nguyen 2020, p. 146). It could be that with election result denialism, we are dealing with echo chambers in which inconvenient truths are systematically rejected, or perhaps with epistemic bubbles within echo chambers. The 2020 election result denialism could even be constructed as an echo chamber within a larger right-wing echo chamber, where insiders have been systematically taught to distrust views that go against the status quo within the echo chamber. At least in the United States, the present-day partisan divisions seem to lend support to such a thought, with echo chambers existing on both ends of the political spectrum.

When deriding established news sources such as *CNN*, *The New York Times* and *The Washington Post*, Trump has frequently used the phrase 'the

fake news media' as an umbrella term to refer to such media outlets (Gelfert 2021). Nguyen (2020, p. 142) argues that the mechanism of systematically discrediting outsiders is like that of indoctrination into a cult, in which members are left "overly dependent on approved inside sources for information". As outside voices are pre-emptively discredited, it is not enough to be exposed to new sources of information, as people have been actively encouraged distrust any evidence or anyone they view as outside sources. This means that the question about culpability of bad beliefs or ignorance becomes even more of a contextual matter. If there is a way out of the grip of an echo chamber, Nguyen (2020) suggests that it must be through repairing trust between members and non-members in some way.

Some have argued that the effects of echo chambers are overestimated. They point out, for example, that most people in the US prefer to watch news on channels without a strong partisan bias (Guess, Nyhan, Lyons and Reifler 2018). Still, even if only a minority of people consume partisan media sources, the impact of these sources can be amplified through inter-personal discussions in the real world when people who consume partisan media discuss the partisan point of view with others, persuading those who do not watch or listen to them (Druckman, Levendusky and McLain 2018). Importantly, it has been suggested that the indirect polarizing effect on the opinions of people who do not consume partisan media, but discuss them with those who do, could be larger than the polarizing effect of consuming the partisan news sources directly (Druckman, Levendusky and McLain 2018). The research suggests that this is most likely due to the social pressure to conform to the common epistemic position in a homogenous group: by going along, you are signaling to others that you belong to the group. This underlines the importance of remembering the effects that our social circles and face-to-face communication have on our epistemic neighborhoods, even in the age of social media.

The distinction between echo chambers and epistemic bubbles has also been challenged. One such criticism is that Nguyen focuses too much on cognitive explanatory mechanisms. When we form beliefs about issues, what matters is not just the background attitudes, credibility assignments, and disagreement-reinforcement mechanisms, but also how we are motivated to take epistemic positions that express our social group memberships (Munroe 2023). Such mechanisms are affective rather than cognitive. Interestingly, it has been suggested that the main drivers behind climate denialism are biases based on group membership (e.g. Carmichael, Brulle & Huxster 2017; Kahan, Peters, Wittlin, et al. 2012). This means that people are already motivated to

look for in-group signals about who to trust and use their social identities in deciding who the experts are (Munroe 2023). On top of important social factors like signaling, it also "feels good" for psychological reasons to engage with news on social media that fits one's existing preferences and biases (Garz, Sörensen & Stone 2020). You get satisfaction for getting affirmation for your beliefs.

Regardless of the exact mechanisms behind them, bad epistemic neighborhoods can overlap and most of us probably occupy several of them. After all, misinformation abounds in our current world, which is both complex and interconnected, and overflowing with information and misinformation alike. If you live in a bad epistemic neighborhood in relation to some normatively important issue, this arguably reduces the culpability for your ignorance related to that matter. It can even make you a victim of these neighborhoods in a way. The prevalence of post-truth narratives, including fake news and bullshitting, has been argued to make people feel confused, disoriented, and powerless (Rietdijk 2024). It has also been suggested that in some cases, the inability to distinguish real experts can excuse false beliefs, and that the individual people who are active disseminators of misinformation can be at the same time the perpetrators and the victims of epistemic injustice (Tappe and Lucas 2022).

There are two predominant strategies for addressing the spread of misinformation: (1) emphasizing individual consumer responsibility, and (2) promoting technological interventions aimed at reducing the cognitive load on individuals, particularly by modifying how information is presented online (Gelfert 2021). Examples of the latter include measures such as automated fact-checking or the labelling of questionable content on social media platforms. However, such technological interventions pose the risk of creating an 'arms race' between the social media platforms and producers of fake news because any viable algorithm designed to detect misinformation can be exploited or circumvented (Gelfert 2021).

Emphasizing individual responsibility neglects the broader context of bad epistemic neighborhoods, focusing instead on measures targeted at individuals, such as enhanced media literacy. This overlooks the fact that reasoning abilities can be employed in self-serving ways: people tend to engage in motivated reasoning and be very good at coming up with justifications for their bad beliefs and actions. In addition, lamenting about the reduction in critical thinking skills is at odds with evidence suggesting that it is older consumers, rather than younger ones, who are more likely to share fake news (Gelfert 2021). Individualistic focus is also problematic for a more general reason: it leaves the picture of epistemic responsibility incomplete (Millar 2021). It also

underestimates the collective phenomenon of 'backfire effects,' where individuals, when confronted with criticism, often become more entrenched in their beliefs (Gelfert 2021). One of the primary forces that leads to the omission of relevant epistemic sources is selective exposure: the tendency to seek like-minded sources (Nguyen 2020). Selective exposure can also be about our social circles, the people whose testimony we trust (Rini 2017). Both ways of ending up in a bad epistemic neighborhood can be unintentional and unmalicious, or both.

Because the question of culpability of individuals within bad epistemic neighborhoods is contextual, my inquiry hereon focuses on the responsibility of the collective agents who impact the epistemic neighborhood.

II. Misbelievers and disinformers

I now turn to the responsibility of collective agents who systemically spread misinformation, distinguishing two motivations for it. I will furthermore propose that these motivations have an impact on their responsibility. While motivations of individual people are often inherently private, and thus cannot be easily publicly verified, this is different with collective agents. In most cases, collective agents must base their actions and assertions on discussions and debates between members about what to do and what to say. Thereby their motivations can be detected from things like their external communications, shareholder or stakeholder meetings, or collective policies. Such documents and actions form a certain narrative about the motivation of the group agent, which means that the motivations of collective agents can often be narratively stipulated from the outside (Hormio 2024). [5] The narrative might be contradictory, of course, but this only means that the position of that group agent is contradictory in relation to that issue.

Internal communications are often more revealing than external, but these too can be made public through leaked reports, whistleblowers, and so on. Naturally, some collective motivations remain a secret, for example due to a secretive corporate culture, but my point is that this is not always the case.

[5] I have argued earlier (Hormio 2024) that we can distinguish honest and mistaken group assertions from those motivated by lying intentions by tracing their narrative roots. This includes looking at if the group statement conflicts with group knowledge on the matter, and if the process of gathering new evidence to form the group position behind the statement is formed in good faith.

Another example is provided by the discovery phase of pretrial procedures, where the parties exchange information and evidence about a case, and may request (among other things) copies of written documents and other internal communications, such as text messages and emails. I will return to this example later in this section.

Although assigning individuated blame is not always helpful when there are many actors involved, especially when the causes are complex and partly systemic, it is still possible to identify some actors that have clearly acted in bad faith. While it does not take away the need to address structural features and to build better epistemic scaffolding in societies, highlighting the role of collective agents in creating bad epistemic neighborhoods locates the discussion at the appropriate level, away from questions relating solely to individual epistemic agents. The collective level is also important because large collective agents can be held to more stringent epistemic standards than individuals. As collective agents like states or corporations have a much larger capacity to process information than individuals, we should not expect the same things epistemically from individuals as we can from collective agents (Vanderheiden 2016). Compared to individual people on the lookout for fallacies, many large collective agents have completely different resources at their disposal to gather the information and expertise necessary to form and express informed views about issues that pertain to their area of operation. This gives them both epistemic power and additional responsibilities, or so I will argue. I should add that even when we cannot access the motivation of a collective agent in some context, the discussion in this paper can still be relevant for political purposes also, such as evaluating the actions of collective agents in the public sphere. This is because it aims to draw up the general outline of how different motivations affect responsibility of collective agents, which is relevant to drawing up both policy and legal measures. The discussion could also provide a basis for collective agents' self-reflection.

Although the agents who create misinformation can be individuals or collectives, I will focus on collective agents. After all, it is exceedingly rare for an individual to be influential enough on their own to affect a large epistemic community without the backing of some collective entity or other that helps them the reach a wide audience, and/or without the power that a role gives them, making them a representative of a collective agent. [6] An epistemically influential individual does not have to have an official role in a collective, but

[6] Individuals can of course go rogue and fail to act within their roles, but I leave this issue aside here.

even then their epistemic platform is usually provided by a collective entity, such as a social media app. Naturally, having an institutional role will significantly amplify an individual's reach. This was the case with Trump: his made-up statements about election fraud gained prominence precisely because of his position as a president.

Deliberately deceptive misinformation can be called *disinformation*. This sub-category of the wider category of misinformation refers to cases in which the agent is deliberately issuing misinformation. To illustrate this difference, consider a new cult that is created solely to make its inventor and leader a rich man. He makes up the story for the cult and slowly starts indoctrinating followers to the story. The story coming from the leader is disinformation, as it is deliberatively misleading. However, that same story coming from a brainwashed cult member is misinformation, since the cult member is genuinely trying to get outsiders to 'see the truth', so there is no deliberate misleading of others involved. I use this distinction to highlight two different kinds of motivations for collective agents who have polluted epistemic neighborhoods.

Misbelievers are engaged in spreading misinformation for ideological or related reasons. The goal of their public communications is to try to convince other epistemic agents to share their ideology and to push back on facts which they cannot accept, for example to deny that the election result was accurate and not a result of widespread voter fraud. They thus want others to become believers in the narrative of a stolen election. Naturally, if they sincerely believe the falsehoods, they will not conceptualize it in such a way themselves.

A possible example of a collective agent falling under the category of a misbeliever in relation to the 2020 election is One America News Network, or One America News (OAN) for short. While it does not have the same audience size as Fox News or other established news channels, it experienced a significant boost in viewership during and after the 2020 presidential election, particularly from the segment of viewers who were not happy with the mainstream media rejecting Trump's narrative about widespread voter fraud (Mitchel 2021). OAN has been described as far-right channel (Robertson 2024, Sneed and Cohen 2023), as it has given ample airtime for conspiracy narratives and other baseless claims circulating in far-right circles (Bode 2020). Among its staff, OAN employs several people known for their far-right views and for subscribing to various conspiracy theories (Breland 2020). After the election, OAN's founder Robert Herring accused the Democrats on social media of cheating in the election (Bode 2020). According to a defamation lawsuit by an employee of Dominion Voting Systems, pro-Trump cable networks like OAN and Newsmax overstepped legal boundaries by promoting

and seemingly endorsing blatant falsehoods (Birkeland 2021). [7] The non-critical support given by OAN to lies like these about the rigging of voting machines, its hiring of several conspiracy theorists, as well as the unsubstantiated claims made by the network's founder about the election being falsified, all indicate bad beliefs about a stolen election motivated by ideological reasons. Of course, it is impossible to confidently make such an assessment without seeing internal communications of the network, or other such materials. What matters for my purposes is that it is plausible that some of the collective agents involved in the election misinformation were in it for ideological reasons, at least for a large part.

Other potential examples of misbelievers can be found when we look at antagonism towards climate science and especially towards its conclusions that urgent and large-scale systemic change is required to tackle the problem. Such misbelieving can be rooted, for example, in a distaste for environmental regulation, which is seen as a threat to the idea of a free market society (Dunlap and McCright 2011; Lewandowsky, Cook & Lloyd 2018). There are many obvious candidates for collective agents that have impacted epistemic communities on climate change. One of the most prominent examples is provided by The Heartland Institute. Their numerous activities include hosting "skeptical" conferences and other events on climate change science, as well as editing, publishing, and promoting reports by their own Nongovernmental International Panel on Climate Change (NIPCC). The latter produces books with misleading titles such as *Why Scientists Disagree about Global Warming*, to coincide with reports by the Intergovernmental Panel on Climate Change (IPCC) and United Nation meetings to discuss climate action. The Heartland Institute distributes its materials in many ways, including targeting groups like policymakers and teachers. Ideology can leave denying or discrediting the scientific consensus as the only cognitive and argumentative option (Lewandowsky, Cook & Lloyd 2018). The Heartland Institute could be an example of this. This kind of psychological denialism can be compared with anti-vaxxers, who spread misinformation because they sincerely believe that vaccines are dangerous. Misbelievers could themselves be products of bad epistemic

[7] Dominion Voting Systems is a company that provides election technology like voting machines and software to various jurisdictions across the United States. These include key swing states such as Georgia, which became a focal point of Trump's anger. Dominion and its employees were subsequently at the centre of various conspiracy narratives. There are several similar defamation lawsuits either pending or settled, spearheaded by Dominion itself or another voting machine company Smartmatic. Both were targeted in the right-wing media in relation to the 2020 presidential election.

neighborhoods. One could speculate, for example, about the role of McCarthyism or Cold War propaganda in response to the threat of the Soviet Union vis-à-vis the distaste that many people in the U.S. have for so-called Big State and regulation.

The second motivation for agents contributing to bad epistemic neighborhoods is more cynical, as the agents know that they are engaging in disinformation. They wield disinformation as a tool for some end, such as gaining power, or concerns over profit. I will call these agents *disinformers*. An example is a local politician who knows that the election results were not tampered with but peddles with disinformation because they think that it appeals to their voters and boosts their party's chances in the next election. The motivation is power, disinformation is just an instrument to get there. Another obvious motivation for disinformation is that there are major business interests at stake. Some well-known examples include the tobacco industry, fossil fuel companies, and the sugar industry (Oreskes & Conway 2010). Here disinformation is utilized as a tool to delay meaningful policy measures that could hurt their bottom-line. Disinformers, like the fossil fuel corporations involved in stalling the progress of UN treaties knew the real state of affairs early on, and acknowledged the truth of climate science internally, but chose disinformation as a strategy to protect their profits in the short term (Cook et al. 2019, Dunlap and McCright 2011, Oreskes & Conway 2010). [8] The sophistication of the misinformation techniques used can affect how robust the bad epistemic neighborhoods are.

An example of a disinformer in relation to the 2020 election result is provided by Fox News, a prominent conservative news channel in the United States. It too has faced defamation lawsuits in response to its coverage of the election and famously paid a $787.5 million to settle a lawsuit brought against it by Dominion (Debusmann 2023). As part of the pre-trial discovery process, Dominion gained access to Fox News' internal communications. Through presenting materials drawn from sources such as internal emails and text messages, the lawsuit details evidence that Fox News knew that the allegations against Dominion were not true. [9] Its own personnel referred to the election fraud claims with terms such as "crazy" and "ludicrous", with at least some of

[8] In doing so, they have utilised the tobacco industry's playbook; an industry that was forced to make concessions only after activists successfully pushed for a framework of distrust towards the Big Tobacco (Derry and Waikar 2008).

[9] Dominion Voting Systems v Fox Corporation, Superior Court of the State of Delaware, Brief in Support of Dominion Motion for Summary Judgment on Liability of Fox News Network, Case No. N21C- 03-257 EMD.

them also clearly understanding how damaging they could be to democracy. Initially the network kept to facts and did not give credence to the stolen election claims. However, this soon changed. The internal communications reveal that although Fox News found no evidence to back up the election fraud claims, the channel was worried about declining viewer figures. In particular, many channel insiders expressed worry over losing the trust of their audience, especially in relation to the damage done to their brand after Fox News (accurately) calling Arizona for Biden. Although top executives like Rupert Murdoch were aware that the network was reporting falsities, they continued doing so for financial reasons: to retain those viewers who were drawn towards the more right-wing context offered by OAN and Newsmax. The lawsuit depicts a news network who recklessly and intentionally disregarded the truth for profits. It served disinformation to its audience for instrumental reasons: to protect its profits.

Of course, it is almost impossible to tell from the outside with a certainty if a collective agent is a misbeliever, or if its actions and statements are mixed with some degree of being a disinformer. While lawsuits that go into discovery can provide fascinating access to the inner workings of an organization, it can be hard to make a clear assessment of when a collective agent knowingly makes false statements. In practice, these motivations might mix. Support for climate change action has been delayed by cynical disinformation campaigns, which were amplified by misinformation efforts driven by ideological reasons. Something similar could arguably be said for the attempt to overthrow the election. There could also be additional motivations, like creating a sense of epistemic safety for members (Furman 2023). A collective agent could also be culpable for displaying other kinds of negligent epistemic behavior, apart from misbelieving or disinforming. For example, epistemic pollution can be created as a side effect of things like wanting to generate more traffic to one's website through promulgating fake news and thereby gaining more profits (Fallis and Mathiesen 2019). Denying facts is not at the center of such behavior. Collective agents can also make epistemic neighborhoods worse without intending to do so. Sometimes bad epistemic neighborhoods are generated as an unforeseen side-effect of some other activity. Some of the most significant examples of epistemic pollution have arguably been unforeseen side effects of new information technology. One example is YouTube's algorithm designed to increase revenue by recommending other videos that viewers might like, duly ending up exposing people susceptible to conspiracy thinking to new conspiracy theories (Landrum, Olshansky & Richards 2021). Still, my focus is on misbelieving or disinforming, because I believe that they represent the two

major motivations for misinformation (even if not necessarily the main sources of it).

More generally, any large-scale changes to our epistemic environment will likely have epistemic costs as well as benefits. With the rapid development of new information technologies, we are often playing catch-up with the broad picture of the impacts. While it is given that social media companies and other such tech companies should work to fix their algorithms once problems arise (and are blameworthy if they fail to do so), the degree to which technology companies should be able to predict the impacts of their innovations before they are launched is an open question that falls outside the scope of this paper. There will also be collective agents who have not done anything to make epistemic neighborhoods worse, but who have the power to make it significantly better and who could arguably have a forward-looking obligation to improve the epistemic neighborhood.

If one bad epistemic neighborhood can indoctrinate you to another, or lowers the threshold for starting another, the original culprits instigating bad epistemic neighborhoods could have done so with "good" intentions. In other words, if we allow for the agent's motivations to impact on whether something is misinformation or disinformation, misbelievers could have engaged in misinformation (albeit without realizing this), rather than disinformation, even when they are creating echo chambers or other mechanisms, perhaps based on affective reasons. Individuals and collectives impact each other and can create their own feedback loops that strengthen certain beliefs. An example would be the local politician trying to appeal to their base by denying the election results. This will be seen by some as further evidence that the election was stolen, thereby increasing support from voters for candidates that deny the results, which can then affect the official party line on the matter (at least locally), again affecting the beliefs of individual voters, and so on.

Be that as it may, epistemological culpability needs to be separated from moral culpability based on epistemic concerns. I will discuss this next, with focus on the responsibility of collective agents who specialize in knowledge.

III. Responsibility of collective agents and consumers of fake news

This section discusses how the motivations of collective agents might affect their culpability for bad epistemic neighborhoods. The two different motivations for misinformation, ideological and cynical, both have bad epistemic

consequences and have contributed to the current political polarization. I will argue that while misbelieving is more culpable in a purely epistemic sense, being a disinformer is more culpable in a moral sense. I furthermore suggest that while collective agents can present us with clear cases of culpability in epistemic matters, when dealing with consumers of fake news and misinformation, we should proceed with a certain level of epistemic humility.

Epistemic institutions present a special case for the responsibility of collective agents. The unifying feature of all epistemic institutions is that knowledge is at the core of their operations (Miller 2022). [10] In what could be called as knowledge institutions (Hormio and Reijula 2024), the purpose of the collective agent centers around knowledge, whether this is acquisition of new expert knowledge or dissemination of knowledge. Examples include universities, intelligence agencies, schools, and think tanks. We hold such collective agents to higher epistemic standards than other collective agents. Media institutions are clearly also epistemic institutions, but we do not hold them quite to the same level of epistemic standards as knowledge institutions. For example, magazines and journals focusing on hobbies or lifestyles are in the business of disseminating knowledge, but their focus is more on practical know-how, information and opinion, not necessarily getting to grips with the truth of the claims. In contrast, the focus of news outlets is more strictly on claims that can be verified. Correspondingly, we tend to hold news outlets to quite high epistemic standards. However, this is not the same standard as for knowledge institutions. The very nature of reporting on breaking news and developing stories necessitates that we should not set the epistemic standard so high that the speaker should be able to guarantee the truth of the claims, although the reporting implicitly contains the idea that the reporter can vouch for their belief that the claims presented are veridical (Gelfert 2021), that is, they coincide with reality.

We rely on institutional expertise in many areas of our daily lives, including in epistemological matters. For democratic social systems to work, we rely on collective agents not to deceive us in matters such as election results, science, healthcare, infrastructure, and numerous other things. As Neil Levy (2022, 94) puts it: “Without heavy-duty social and environmental scaffolding, even virtuous agents can’t reliably acquire knowledge about difficult and complex issues”. With many important issues, such scaffolding can be patchy at best, with disinformation lurking even within prima facie respectable epis-

[10] According to Miller (2022), social media providers are not so much epistemic institutions as disseminating institutions.

temic sources like news channels and speeches by politicians. Institutions within democratic systems arguably have an obligation not to deceive us, especially when they have more information about an important issue than we do and they are viewed as trustworthy sources of information, but as we have seen, some of them do so regardless.

Seana Shiffrin (2014) has argued that the primary wrong in lying is that it abuses the mechanism by which we provide reliable testimonial warrants to each other. [11] We need this mechanism to cooperate with each other, so the distinctive wrong in lying is the abusing of this mechanism. In other words, she underscores undermining testimonial trust as the main wrong in lying, rather than the deceptive effects that the lie might have on its audience. According to Shiffrin, testimonial practices based on sincere speech are "fundamental components of a social environment that supports the moral agency of thinkers" (2014, p. 117), and damage to these testimonial practices also jeopardizes the rational basis we have for supporting them. After all, if lying becomes common practice, it erodes social trust and makes cooperation increasingly difficult or even unsustainable. The election denialism and the Capitol attack showed how harmful false beliefs can be and how bad epistemic neighborhoods can erode social cohesion. This trust in the testimony of others is more fragile when what they tell us poses a threat to our existing ways of living. Therefore, attacks on things like the scientific consensus on an important matter should be made in good faith: only when you actually believe what you are claiming.

Am I then suggesting that attacks on election results or on climate science by misbelievers are somehow less harmful than those made by disinformers? Not really: both form of denialism can be equally polluting, regardless of if the intention is to be misleading or not. If false claims become standard practice in news, we can no longer trust the news. Misbelieving could even be more harmful than being a disinformer, because ideology can make messages more potent within an already ideologically susceptible epistemic neighborhood.

Although I have focused on the false beliefs of people within an epistemic neighborhood, the pollution also has other harmful effects. For example, by refusing to participate in an open debate about their claims, the polluting agents are putting obstacles in the path of a genuinely pluralistic dialogue through repeating already dismissed claims (Shaw 2021). Moreover, this is

[11] This argument has Kantian roots, as it aligns with Kant's emphasis on truthfulness as a fundamental duty in communication and the preservation of trust in social interactions.

linked to an important phenomenon, affective polarization, which refers to emotional division between political groups and leads to a rise in mistrust toward the political outgroup. This type of mistrust is primarily driven by emotions and feelings rather than grounded in a rational, shared understanding, or a logical narrative that explains why members of the opposing party should not be trusted (Munroe 2023). In this case, mistrust is based on personal aversion to the outgroup rather than on substantive disagreements or evidence-based critiques of their actions or policies. While bad beliefs and harmful narratives can enforce the phenomenon, it is a step away from the kind of reasons we are used to looking at in epistemology.

Still, the upshot of my argument is that collective misbelievers are less culpable moral agents than disinformers, if they are honestly mistaken about the beliefs that they hold to be true. My suggestion is that while the moral culpability of misbelievers might be reduced to some degree by their motivation, as epistemic agents, collective misbelievers are clearly culpable for their bad beliefs. Due to their greater epistemic capacities, the epistemic culpability of collective agents is often a clearer issue than with individual agents. This holds even if their members are indoctrinated by bad epistemic neighborhoods. With the overwhelming consensus among climate scientists, it would be hard to argue that misbelievers like The Heartland Institute and others are engaging with available evidence in good faith. Its members and supporters might be ideologically motivated by wanting the science not to be true, as the need for climate change regulation goes against the stated goal of the institute to defend limited government. But as collective agent specializing in information (a think tank), they are being an epistemically irresponsible knowledge institution if their output is clouded by their ideology. The same arguably goes for OAN and other collective agents who disregard the facts over affect in their output, even when they present themselves as a news outlet.

Fox News as a disinformer is more careful as an epistemic agent when forming its beliefs, compared to a misbeliever, as it has taken the facts seriously when looking into the matter. After all, it clearly knew that the election fraud claims were false, and chose not to report on them initially. However, by being intentionally deceptive and engaging actively in disinformation, it is also arguably a more culpable agent in moral terms. It clearly should have known better. Fox News knew what it was involved in and was not under any ideological delusions, like saving democracy from some conspiracy or another. This example shows that careful epistemic agents might nonetheless be morally culpable agents on epistemic grounds. In other words, epistemic blame and moral blame based on epistemic reasons can come apart.

If you recall, Fox News was worried about losing viewers and felt like they should cater to their audience by giving airspace and therefore credibility to false claims about a rigged election. Fox News arguably saw themselves as catering to a customer demand. The question arises: should consumers of fake news and misinformation bear some of the burden for their bad epistemic habits (Croce and Piazza 2021)? I argued earlier than within many bad epistemic neighborhoods, an individualistic starting point for such questioning is not particularly fruitful. However, we can look at the issue from the point of view of consumers of fake information as a group. The group will not be a group agent, but rather an unstructured group whose constituents share a normatively important feature: they consume a lot of fake news and misinformation. Do unstructured groups such as far-right Fox News and OAN viewers share responsibility for supporting a news outlet that peddles in misinformation?

The relationship between individuals and collectives is complicated. Not only do collective agents comprise of individuals-in-roles, but they are also made possible by individuals. These individuals are not just the powerful leaders of the collective agents, but also their customers and other supporters. Take The Heartland Institute as an example. According to their website, approximately 70% of their funding comes from individuals. [12] This gives rise to an interesting question about the extent to which individuals donating to the institute are responsible for creating misinformation. In broader terms, consumers of fake information and news could be culpable of creating a demand for misinformation. After all, agents within a bad epistemic neighborhood can both consume misinformation and contribute to creating it. This further complicates matters for individual responsibility as many of these people might epistemically buy into denialism, so they are not contributing to deceiving others intentionally. Rather, they could simply be mistaken; it is possible to sincerely believe in a falsehood, especially if you are deeply in thrall to some ideology. This is why it seems bit too demanding to argue that citizens in democratic states have a shared responsibility to counter false information, and that citizenry can reasonably be expected to take steps towards restricting the spread of false claims (like Millar 2021 does). Some sections of the citizenry, certainly, but also those deeply in thrall to ideology and misinformation? It does not sound feasible for the kinds of reasons discussed earlier in relation to individual culpability. At the same time, it seems that we want to

[12] According to www.heartland.org (accessed 14 April 2021), in 2017 the think tank had approximately 5,000 supporters, with 70% of its income coming from individuals, 22% from foundations, and 6% from corporations.

describe some kind of shared responsibility to those who actively seek out false narratives, or who fund think tanks that work to discredit science. Even if this is so, it seems disingenuous to put the blame on consumers alone, or even for the most part, as consumer demand is often created and sustained by collective agents. [13] Yet we do not want to let all the individuals of the hook, especially those who fund or otherwise enable misinformation for malicious motivations, like racism or misogyny. Be that as it may, such a responsibility debate will have to be contextual.

Populist rhetoric can make it harder to contain fake news, as it makes a sharp division between the people and the elite, who is allegedly corrupt. Even so, I think that liberal democratic governments must be careful on relying more on experts' advice to counterbalance denialism and other epistemic failures among voters. A shift towards more epistocratic governance is only likely to exaggerate the underlying problems, such as lack of trust in institutions, in addition to raising questions about justification. Narrow technocratic framings can also lead us to look for solutions in the wrong places.

I find that a better response is a certain level of epistemic humility. In our polarized world, it might be what we need in our daily interactions with other individual epistemic agents. The kind of epistemic humility I have in mind is akin to intellectual humility, which encourages one to acknowledge the boundaries of one's knowledge and accept the possibility of being mistaken. It involves an understanding that one's beliefs, perspectives, or assumptions may be flawed or incomplete. Once we adopt such an attitude, it fosters a willingness to listen to others and to re-evaluate one's own views. I like the idea presented by Elisabetta Galeotti and Federica Liveriero (manuscript) that intellectual humility is similar to moral responsibility that is associated with one's social and epistemic privileges (rather than just an epistemic obligation tied to a fallibilist approach to knowledge). Because of their privileged position, reasonable individuals bear the responsibility of addressing social imbalances when it comes to epistemic matters, such as misinformation and conspiracies. In other words, the initial steps toward rebuilding mutual trust and fostering a cooperative democratic environment lie with those who are capable of adopt-

[13] Examples of demand creation include mobile phone companies introducing new models to get consumers to update to the latest model even when their existing phones are still working, or fashion brands pushing for ever-shorter cycles on clothing trends. Something similar could be argued to apply to misinformation, for example through the way in which social media algorithms tend to favour high emotion polarising content, including conspiracy theories, and click-bait journalism.

ing such an attitude—specifically, individuals whose belief systems are not anchored in personal commitments or identities.

This does not mean that we should adopt a similar attitude towards collective agents, though. Collective agents, especially epistemic institutions, such as news channels, can and should be held to more stringent epistemic standards than individuals. Although news does not claim to disseminate knowledge, they should mirror what the reporter believes to coincide with reality. With misbelievers, misinformation could be based on bad beliefs that the reporters and the head of the network themselves have. Such news channels fare poorly as epistemic agents. In real life, this can be reflected in market-based consequences, such as marginal viewer figures or being dropped by cable channel providers, both of which have happened to OAN. Policy measures to prevent such channels from presenting themselves as news outlets could also be needed.

Disinformers are often more careful epistemic agents, as they know that what they are claiming is false. However, they are also morally more culpable agents, as they place profits or other such motivations over honesty. Such a stance is especially harmful when the collective agent is an epistemic institution. Fox News was able to give credibility to baseless claims precisely because it is viewed as a reputable news organization based on its past performance. It still employs ambitious journalists who subscribe to high standards in their reporting. The problem is that by attaching to the same channel, the more far-right views got the kind of legitimacy that OAN or other epistemically inferior collective agents would never have been able to give them. Fox News chose to trade its credibility for money, giving a powerful platform to dangerous accusations that eroded public trust in democracy. There should be measures to penalize such actions by a news organization through means other than just by going through the courts to bring on defamation lawsuits. [14] Some possible measures include critically evaluating broadcasting licenses of news outlets engaged in disinformation, but such measures should be utilized only through extreme care. Freedom of speech is paramount for democracy and for media actors to be able to act as independent watchdogs for the government within democracies. In a highly polarized landscape, like the United States at the moment, such measures could easily backfire and be taken to extremes. It is important to keep in mind that a balanced democracy needs conservative and progressive voices both, although no democracy needs poorly run news organizations.

[14] Defamation law has a limited capacity to bring justice to the public for misinformation, especially as the agreements lack disclosure and accountability (Levine 2024).

A big issue looming in the background is that sometimes bad epistemic neighborhoods are not the root-causes of bad beliefs. Rather, they are the symptoms of something else that has gone wrong in a society. Bad epistemic neighborhoods can certainly amplify and harden polarization, but they are not necessarily the cause of it. Other large-scale trends, such as widening inequality of opportunities, could be driving the polarization. The rapid rise in the availability of AI technologies, coupled with social media algorithms that favor items that raise the pulse of its audience, means that convincing misinformation is easier and less costly to manufacture than ever before, even by isolated individual agents, making it easier for fringe views to gain wider audience. Such wider trends need to be tackled in tandem with measures that target bad epistemic neighborhoods.

Conclusion

The collective agents that create and amplify bad epistemic neighborhoods can be divided roughly into those that do so mostly for ideological reasons (*misbelievers*), and those that do so for instrumental reasons (*disinformers*), for example to gain more power or monetary profit. These two motivations impact the responsibility of the collective agents that help to create bad epistemic neighborhoods. While a disinformer of facts might be more careful as an epistemic agent than a misbeliever when forming its beliefs, by being intentionally deceptive and engaging actively in disinformation, it is also arguably a more culpable agent in moral terms. Furthermore, I have argued that even if there is some degree of blame for the individuals who holds bad beliefs or are ignorant of normatively important matters, the blame is better directed at the collective agents creating and contributing to bad epistemic neighborhoods, such as media organizations and lobby groups. More generally, we should not be too quick to judge and ridicule the views of our fellow citizens, as that often only leads to further polarization. We should instead aim for a certain level of epistemic humility: we all have some bad beliefs, and we all occupy some bad epistemic neighborhoods. This does not mean that bad beliefs are harmless or that they should not be criticized. They should, but the criticism should be directed more towards the collective level, and our energies are better spent when focused on improving the epistemic scaffolding in our societies.

Works Cited

Arpaly, N. 2002. *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford University Press.

Birkeland, B. 2011. "Newsmax Issues Retraction and Apology to Dominion Employee Over Election Stories." *NPR*, April 30, 2021. www.npr.org.

Bode, K. 2020. "OAN Is So Dangerous Because It Looks Like a Real News Channel." *VICE*, November 25, 2020. www.vice.com.

Breland, A. 2020. "Meet the Propagandists and Conspiracy Theorists Behind the One America News Network." *Mother Jones*, June 9, 2020. www.motherjones.com.

Carmichael, J.T., R.J. Brulle, and J.K. Huxster. 2017. "The Great Divide: Understanding the Role of Media and Other Drivers of the Partisan Divide in Public Concern Over Climate Change in the USA, 2001–2014." *Climatic Change* 141(4): 599-612.

Cook, J., G. Supran, S. Lewandowsky, N. Oreskes, and E. Maibach. 2019. *America Misled: How the Fossil Fuel Industry Deliberately Misled Americans About Climate Change*. George Mason University Center for Climate Change Communication.

Debusmann, B. Jr. 2023. "Fox News Settles Dominion Defamation Case for $787.5m." *BBC*, April 19, 2023. www.bbc.com.

Derry, R., and S.V. Waikar. 2008. "Frames and Filters." *Business & Society* 47(1): 102-139.

Dominion Voting Systems v Fox Corporation, Superior Court of the State of Delaware, Brief in Support of Dominion Motion for Summary Judgment on Liability of Fox News Network, Case No. N21C-03-257 EMD.

Druckman, J.N., M.S. Levendusky, and A. McLain. 2018. "No Need to Watch: How the Effects of Partisan Media Can Spread via Interpersonal Discussions." *American Journal of Political Science* 62(1): 99-112.

Dunlap, R.E., and A.M. McCright. 2011. "Organized Climate Change Denial." In *The Oxford Handbook of Climate Change and Society*, edited by J.S. Dryzek, R. B. Norgaard, and D. Schlosberg, 144-160. Oxford University Press.

Fallis, D., and K. Mathiesen. 2019. "Fake News is Counterfeit News." *Inquiry*, 1-20.

Furman, K. 2023. "Epistemic Bunkers." *Social Epistemology* 37(2): 197-207.

Galeotti, E., and F. Liveriero. (manuscript). "Preposterous Fake News, the Breach of Democratic Trust and the Civic Virtue of Reasonableness."

Garz, M., J. Sörensen, and D.F. Stone. 2020. "Partisan Selective Engagement: Evidence from Facebook." *Journal of Economic Behavior and Organization* 177: 91-108.

Guess, A., B. Nyhan, B. Lyons, and J. Reifler. 2018. "Avoiding the Echo Chamber About Echo Chambers: Why Selective Exposure to Like-Minded Political News is Less Prevalent Than You Think." *Knight Foundation White Paper*. John S. and James L. Knight Foundation.

Gelfert, A. 2021. "What is Fake News?" In *The Routledge Handbook of Political Epistemology*, edited by M. Hannon and J. de Ridder, 1st ed. Routledge.

Haack, S. 1997. "'The Ethics of Belief' Reconsidered." In *The Philosophy of Roderick Chisholm*, edited by L.E. Hahn, 129-144. Open Court.
Hormio, S. 2024. "Group Lies and the Narrative Constraint." *Episteme* 21(2): 478-497.
Hormio, S., and S. Reijula. 2024. "Universities as Anarchic Knowledge Institutions." *Social Epistemology* 38(2): 119-134.
Jolley, D., and K.M. Douglas. 2014. "The Social Consequences of Conspiracism: Exposure to Conspiracy Theories Decreases Intentions to Engage in Politics and to Reduce One's Carbon Footprint." *British Journal of Psychology* 105: 35-56.
Kahan, D.M., E. Peters, M. Wittlin, et al. 2012. "The Polarizing Impact of Science Literacy and Numeracy on Perceived Climate Change Risks." *Nature Climate Change* 2: 732-735.
Landrum, A.R., A. Olshansky, and O. Richards. 2021. "Differential Susceptibility to Misleading Flat Earth Arguments on YouTube." *Media Psychology* 24(1): 136-165.
Levine, S. 2024. "What Price Are US Media Outlets Paying for Spreading Election Lies?" *The Guardian*, September 30, 2024. www.theguardian.com.
Lewandowsky, S., J. Cook, and E. Lloyd. 2018. "The 'Alice in Wonderland' Mechanics of the Rejection of (Climate) Science: Simulating Coherence by Conspiracism." *Synthese* 195: 175-196.
Levy, N. 2022. *Bad Beliefs: Why They Happen to Good People*. Oxford University Press.
Mason, E. 2015. "Moral Ignorance and Blameworthiness." *Philosophical Studies* 172: 3037-3057.
Millar, B. 2021. "Shared Epistemic Responsibility." *Episteme* 18(4): 493-506.
Miller, S. 2022. "Epistemic Institutions: A Joint Epistemic Action-Based Account." *Philosophical Issues* 32: 398-416.
Mitchell, A. 2021. "Large Majorities of Newsmax and OAN News Consumers Also Go to Fox News." *Pew Research Center*, March 23, 2021.
Munroe, W. 2023. "Echo Chambers, Polarization, and 'Post-Truth': In Search of a Connection." *Philosophical Psychology*, 1-32.
Nguyen, C.T. 2020. "Echo Chambers and Epistemic Bubbles." *Episteme* 17(2): 141-161.
Nottelmann, N., and P. Fessenbecker. 2019. "Honesty and Inquiry: W.K. Clifford's Ethics of Belief." *British Journal for the History of Philosophy* 28(4): 1-22.
O'Connor, C., and J.O. Weatherall. 2019. *The Misinformation Age: How False Beliefs Spread*. Yale University Press.
Oreskes, N., and E.M. Conway. 2010. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Press.
Piovarchy, A. 2021. "What Do We Want from a Theory of Epistemic Blame?" *Australasian Journal of Philosophy* 99(4): 791-805.
Rietdijk, N. 2024. "Post-Truth Politics and Collective Gaslighting." *Episteme* 21(1): 229-245.

Rini, R. 2017. “Fake News and Partisan Epistemology.” *Kennedy Institute of Ethics Journal* 27(S2): E43-64.
Robertson, K. 2024. “Smartmatic and OAN Settle Defamation Suit.” *The New York Times*, April 16, 2024.
Robichaud, P. 2017. “Is Ignorance of Climate Change Culpable?” *Science and Engineering Ethics* 23: 1409-1430.
Ryan, S. 2018. “Epistemic Environmentalism.” *Journal of Philosophical Research* 43: 97-112.
Shaw, J. 2021. “Feyerabend and Manufactured Disagreement: Reflections on Expertise, Consensus, and Science Policy.” *Synthese* 198: 6053-6084.
Shiffrin, S. 2014. *Speech Matters*. Princeton University Press.
Singer, J. W. 2009. “Normative Methods for Lawyers.” *UCLA Law Review* 56: 899-982.
Smith, H. 1983. “Culpable Ignorance.” *The Philosophical Review* 92(4): 543-571.
Sneed, T., and M. Cohen. 2023. “Far-Right Network OAN Settles 2020 Election Defamation Suit Brought by Ex-Dominion Executive.” *CNN*, September 5, 2023. edition.cnn.com.
Tappe, L., and D. Lucas. 2022. “Of Sheeple and People: Echo Chambers, Pseudo-Experts and the Corona Crisis.” *Disputatio: Philosophical Research Bulletin* 11(20): 119-131.
UNITED STATES OF AMERICA v. DONALD J. TRUMP, Criminal No. 23-cr-257 (TSC).
Vanderheiden, S. 2016. “The Obligation to Know: Information and the Burdens of Citizenship.” *Ethical Theory and Moral Practice* 19(2): 297-311.
Wolf, S. 1987. “Sanity and the Metaphysics of Responsibility.” In *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*, edited by F. D. Schoeman, 46-62. Cambridge University Press.

DOI 10.17473/2240-7987-2024-2-11

The Unequal Damages of Fake News: Amplifying Epistemic Inequality and Oppression

Laura Santi Amantini*

Abstract

This paper argues that the harm of fake news lies partly in its inequality-reinforcing effect. It shows how fake news interacts with two background inequalities, namely the unequal distribution of opportunities to become competent knowers and the unequal social status of oppressed minorities. First, I argue that fake news reinforces epistemic inequality among its recipients. The epistemic harm of believing fake news does not uniformly affect all members of society. Instead, epistemic inequality makes some citizens more vulnerable to believing fake news. In turn, believing fake news amplifies epistemic inequality by making those who fall for fake news increasingly less competent as knowers. Second, I argue that fake news amplifies status inequality within society by downgrading the status of minorities who are negatively represented in fake news stories. A significant subset of fake news targets minorities who are not recognized as equal within society, such as immigrants, Muslims and Roma minorities. The harm of being negatively portrayed in anti-minority fake news stories falls asymmetrically on members of the targeted minorities, reinforcing their oppression. Focusing on anti-immigrant fake news as the most widespread and paradigmatic example of anti-minority fake news, I show how it amplifies cultural imperialism, marginalization and systemic violence against members of immigrant minorities. Anti-minority fake news is also specifically harmful to members of targeted minorities as political agents because it contributes to denying their equal status as members of the democratic political community. This subset of fake news stories is therefore particularly corrosive to the democratic system and its core values.

Summary: Introduction. – I. Conceptualizing fake news and its harms for liberal democracies. – II. Citizens' unequal vulnerability to fake news and the amplification of epistemic inequality. – III. Oppressed minorities' unequal vulnerability to be targets of fake news. – IV. Amplifying oppression: the case of anti-immigrant fake news. – Conclusion. – Works Cited.

* ORCID: 0000-0001-5604-9389.

Introduction

Pope Francis endorsed Donald Trump in the United States presidential elections. Hillary Clinton sold weapons to ISIS. These are famous fake news dating back to 2016. The electoral campaign for the 2016 presidential elections in the United States (US) was characterized by an unprecedented proliferation of fake news on social media, which attracted public and scholarly attention. A growing literature has developed across political theory and social epistemology on the nature of fake news and its dangerous effects on democratic societies. It has been argued that the spread of fake news is harmful to citizens as knowers given that it can affect their beliefs and epistemic capacities, even inadvertently (see Levy 2017; Brown 2021). Furthermore, from a democratic angle, it has been argued that fake news is particularly harmful to both citizens and democratic institutions. Indeed, it affects the political agency of citizens. The perception of fake news being widespread lowers their trust in fellow citizens, in the value of democratic procedures and their outcomes, and undermines the constitutive values of democracy. Ultimately, widespread fake news could put at risk the stability of democratic institutions (see Rini 2021; Reglitz 2022). However, the philosophical literature on fake news has not engaged much with the interplay between fake news and background social injustices among citizens.

This paper defends the claim that part of the damage of fake news consists in its inequality-amplifying effect. Fake news amplifies epistemic inequality among its recipients. It also amplifies status inequality within society by downgrading the status of minorities who are negatively represented in fake news stories. Firstly, I argue that the harm of fake news on citizens as knowers falls primarily on those citizens who are already disadvantaged in terms of epistemic skills. Indeed, the epistemic harm of believing fake news does not uniformly affect all members of society. Instead, it is asymmetric since it is mediated by pre-existing epistemic inequality and contributes to deepening such inequality, further eroding the epistemic capacities of those who are more vulnerable to falling for fake news. Secondly, I argue that the content of fake news matters in determining its unequal damages. Several famous fake news stories target specific individuals belonging to the political elite (e.g., Hillary Clinton). Yet many others target members of minorities who are routinely discriminated against and marginalized in Western societies because of their ethnic background, immigrant origin, religious affiliation, and so forth. The harm of being negatively depicted in anti-minorities fake news asymmetrically falls on members of targeted minorities, strengthening their oppression. Anti-minority fake

news is also specifically harmful to members of targeted minorities as political agents because it contributes to denying their equal status as members of the democratic political community. Hence, this subset of fake news is particularly corrosive to the democratic system and to its basic values. Briefly, the diffusion of fake news, whatever its content, is distinctively harmful to epistemically disadvantaged recipients, while anti-minority fake news is specifically harmful to oppressed minorities. Such unequally distributed damages of fake news are normatively salient since they contribute to worsening social injustice and hindering political equality among citizens of democratic societies.

There are at least two important advantages in bringing into light the unequal impact of fake news on citizens of democratic societies caused by its interplay with pre-existing inequalities. Theoretically, it allows us to develop a more realistic and refined account of what is wrong with fake news. Normatively, it has important implications for how democratic institutions should respond to fake news. When theorizing what is wrong with the spread of fake news and how liberal democracies should address it, we should consider that oppressed minorities are distinctively targeted by this new form of misinformation and that anti-minority fake news specifically damages members of targeted minorities by reinforcing their oppression. When evaluating the role of citizens themselves in spreading misinformation, we should take into account the sources of unequal epistemic inequality that make some citizens more vulnerable than others to believing fake news. Moreover, we should consider that believing fake news amplifies such an epistemic inequality. Hence, addressing the unequal damages of fake news would require tackling structural causes of epistemic vulnerability. Taking background epistemic inequality into account also suggests not overstating the individual responsibility of citizens in spreading fake news and avoiding attributing equal blame on all those who spread it for their failure to behave as virtuous epistemic agents. Exploring policy implications falls outside the scope of this paper. The aim, instead, is to offer a more complete picture of the damages of fake news on democratic societies by highlighting how this new form of misinformation interacts with two background inequalities, namely the unequal distribution of opportunities to become competent knowers and the unequal social status of oppressed minorities.

The paper is structured in four sections. Before unfolding my argument, in Section 1, I define fake news, describe it as a new form of misinformation, and present some damages of fake news to citizens and democratic institutions which might seem to apply uniformly on society. In section 2, I argue that the epistemic harm of fake news to citizens as knowers affects citizens unequally. I draw on empirical literature to show that epistemic inequality makes some

citizens more vulnerable to believing fake news and argue that believing fake news amplifies such epistemic inequality, making those who fall for fake news increasingly less competent as knowers. Sections 3 and 4 argue that the content of fake news is relevant in determining what makes fake news specifically harmful to some groups of citizens. In section 3, I show that a significant subset of fake news, rather than targeting political elites, targets those minorities who are not recognized as equals within society, such as immigrants, Muslims and Roma minorities. In section 4, I focus on anti-immigrant fake news as the most widespread and paradigmatic example of anti-minority fake news. Drawing on the theory of oppression offered by I. M. Young's (1990), I argue that anti-immigrant fake news stories amplify cultural imperialism, marginalization and systemic violence against members of immigrant minorities. Therefore, these stories produce a distinctive status harm to members of the targeted minority and are particularly corrosive to the democratic value of equality.

I. Conceptualizing fake news and its harms for liberal democracies

Political philosophy has long dealt with phenomena of disinformation, such as propaganda. Disinformation, conceived as the deliberate spread of false, biased, or misleading information intended to deceive the receiver, can be distinguished from misinformation, which may not involve the deliberate intention to mislead or deceive (Brown 2021). Neither disinformation nor misinformation are new. However, over the last few years, fake news has emerged as a distinct phenomenon. Although the definition is disputed, it can be described as a story that "purports to describe events in the real world, typically by mimicking the conventions of traditional media reportage" (Galeotti and Meini 2022). To be "fake", news need not be literally false: it can be based on insufficient evidence or convey false information through conventional or conversational implicatures. For instance, fake news may refer to real events and yet pragmatically imply an unjustified causal relationship (Galeotti and Meini 2022; Jaster and Lanius 2018). Thus, it differs from genuine mistakes and oversimplifications made in good faith. Some scholars consider deliberate deception as a necessary condition for a piece of news to count as fake: according to Rini (2017, E45), fake news is "known by its creators to be significantly false and is transmitted with the two goals of being widely retransmitted and of deceiving at least some of its audience". Others have challenged this definition, pointing to the case of clickbait articles devised to attract views rather than to persuade readers to believe what they read. Fake

news creators may thus be simply epistemically indifferent and motivated by non-epistemic goals (Jaster and Lianus 2018; Croce and Piazza 2021). Briefly, fake news provides deceitful information, either intentionally or due to a lack of interest in its truth on the part of its creators and disseminators. Even when produced with a deliberate deceiving intention, fake news may well be further disseminated by receivers who genuinely believe it or do not deliberately intend to deceive. Hence, I consider here fake news as a form of misinformation.

Unlike previous forms of misinformation, fake news is created to be spread on the internet, particularly on social media. Social media are characterized by disintermediation (Szakács and Bognár 2021, 8). Indeed, every user can post news, share news, or increase news' popularity by reacting to a post. The lack of gatekeeping functions that journalists and editors play in traditional media facilitates the publication of false, biased, and misleading information. Moreover, social media allow an unprecedented spreadibility of such information. Social media algorithms privilege the spread of content rapidly attracting reactions. Sensational, emotion-arousing fake news can thus be disseminated extremely quickly and widely. Thus, fake news on social media circulates even more quickly and widely than genuine news. Moreover, social media posts can also be easily shared as private messages through instant messaging systems, such as Whatsapp and Telegram, which amplify the diffusion of fake news. The use of these technologies and the active role of social media and instant messaging users as agents of peer-to-peer misinformation make fake news a specific form of misinformation.

Note that social media were not created as information services but as online networks enabling people to interact (Chambers 2021, 151). Social media are used mainly as a form of entertainment requiring limited cognitive effort. Checking sources and questioning the credibility of the information means going beyond the default predisposition to accept the information one receives (Lewandowski 2012). It requires an extra cognitive effort that social media as entertaining environments disincentivize. [1] Moreover, it also requires specific epistemic skills. Indeed, social media blur the boundary between information and entertainment, news and jokes. Navigating the contents of social media thus requires a kind of training which is not needed, for instance, when reading a newspaper. Consequently, as I argue in the next section, citizens who are less epistemically skilled and trained to discern the types of content they encounter on social media are more vulnerable to falling prey to fake news.

[1] As Humprecht (2019) reports, a 2014 study found that roughly 6 in 10 Americans do not even read the articles but only headlines when they come across news on social media.

Political philosophers have already highlighted that the spread of fake news has harmful effects on citizens and on the institutions of liberal democracies. Let us consider, first, the harmful effects of fake news on citizens as fake news consumers. Fake news undermines their personal autonomy by hindering their capacity to appreciate reasons in favor or against a certain position (Brown 2021). Citizens who believe fake news are harmed in their capacities as knowers and agents since they make their judgements based on bad beliefs and false information. What is more, fake news does not have pernicious effects only on those citizens who mistake it as genuine news and make their political judgments accordingly. Levy (2017) argued that even "sophisticated consumers", who consume fake news for fun or to know what credulous fellow citizens believe, can be inadvertently influenced. As Levy puts it, even for those who confidently identify a claim as false, familiarity influences the processing of semantically related claims, contributing to implicit biases. From a democratic perspective, it is particularly relevant that fake news negatively affects citizens in their capacity as political agents, influencing their beliefs, political judgements and electoral choices. This can also have negative consequences in terms of the outcomes of democratic procedures. If one assumes that democracies require informed voters to function well as a collective choice mechanism, the fact that fake news can deceive citizens and even inadvertently influence those who recognize it as false means that fake news damages the outcomes of democratic decision-making.

One may observe that the impact of fake news on citizens' beliefs and its consequences on voting should not be exaggerated. Sharing fake news on social media, or even expressing endorsement for it, does not necessarily imply believing its content. The goal can be just to signal which side one assumes in a polarized political environment. According to Hannon (2021), citizens often declare to believe a given assertion out of "political cheerleading" but may not, in fact, believe it to be true. For instance, Republicans sharing fake news on Hillary Clinton could simply intend to symbolically show their political affiliation and support for Donald Trump rather than express genuine belief in a fact. If Hannon is right, it is not the case that all those who appear to endorse a piece of fake news believe it and behave according to that belief. The disruptive effect of fake news on electoral choices may not be as severe as initially thought. Nevertheless, fake news remains epistemically harmful for those who believe it, however numerous they are.

Furthermore, some have argued that the real threat of fake news does not reside in its persuasive effect on people's beliefs, but in the erosion of epistemic trust among citizens, which in turn undermines the perceived legitimacy

of democratic institutions (Rini 2021; Reglitz 2022). Fake news can be deleterious for democratic institutions regardless of whether it is widely believed or widely influential: what counts is the perception of its being widespread. Citizens of liberal democracies are increasingly aware of the fact that the information they receive can be fake news and that other citizens may believe fake news and pass it on. The awareness of widespread fake news disrupts the usual norms of testimonial knowledge and undermines the epistemic trust among citizens. The perception of other citizens as untrustworthy epistemic vectors, as Rini (2021: 43) notes, induces the belief that those people are untrustworthy as parties with whom they can "negotiate mutual self-government in good faith". Reglitz (2022) makes the implications for the legitimacy of democratic institutions even clearer. If a citizen believes their counterparts not to be competent enough to distinguish facts from lies, or reliable from untrustworthy information sources, this citizen is likely to be doubtful of the reasonableness of their counterparts' political views. Hence, they may judge laws and policies produced by democratic institutions as no longer worthy of their respect, to the extent that they are based on false beliefs and bad choices taken by incompetent and manipulated citizens (see Reglitz 2022, 171, 174). Thus, the perception of widespread fake news, irrespective of whether it is in fact widely believed, undermines the perceived legitimacy of democratic institutions.

Since it does not sufficiently engage with background social injustices, this literature on the harmful effects of fake news might give us the impression that fake news disrupts otherwise functioning democracies. Unfortunately, this is not the case. Existing liberal democracies are not just societies that live up to their democratic values, where citizens have equal opportunities to become competent knowers and treat one another as moral equals. On the contrary, fake news spreads in deeply imperfect democracies and has an asymmetric harmful impact on citizens as receivers and as targets of misinformation, amplifying background social injustices. In the following section, I discuss how epistemic disadvantage makes some citizens more vulnerable to falling prey to online misinformation and how, in turn, believing fake news deepens such an epistemic disadvantage. Next, in sections 3 and 4, I consider how a particular subset of fake news - i.e., anti-minorities fake news - exploits and deepens the oppression of targeted minorities, focusing on anti-immigrant fake news stories as the most prominent and widespread examples of anti-minority fake news.

II. Citizens' unequal vulnerability to fake news and the amplification of epistemic inequality

Epistemic inequality mediates the damage of fake news on citizens as knowers. It makes some of them more vulnerable to believing fake news. Furthermore, it makes believing fake news more detrimental for those who are already epistemically disadvantaged. Someone may immediately observe that exposure to fake news can be harmful to anyone and that surely nobody is immune from falling for fake news. Indeed, humans are all susceptible to cognitive distortions and biases that may lead to attributing unwarranted credibility to a piece of information. Confirmation bias encourages us to take information as true when it confirms previous beliefs, including stereotypes and prejudices. Moreover, fake news on social media and instant messaging systems is typically received from friends and acquaintances, which may grant it unwarranted credibility. Indeed, fake news exploits the mechanism of testimonial knowledge, whereby a person believes a given proposition because the proposition was presented to them by another person. Rini (2017) notes that testimonial norms are ambiguous on social media: sharing a post on Facebook or retweeting on Twitter does not unequivocally mean endorsing it or being sufficiently competent on the matter. Yet, users tend to take information as true, or at least possibly true, when it is passed on from their social media contacts, even when the reported fact sounds weird. Roughly put, the fact that a person we trust endorses P (or might have endorsed P, given that they reported P) is taken as testimonial evidence that P is true (or, at least, might be true).

From the fact that nobody is immune from falling for fake news, however, it does not follow that citizens are equally susceptible to believing fake news. Indeed, citizens do not have equal opportunities to become competent knowers: epistemic goods such as education are unevenly distributed in society and citizens do not develop epistemic skills to the same degree. I argue that epistemic inequality makes some citizens more vulnerable to believing fake news or accepting it as plausible and passing it on, thereby misinforming others. By epistemic inequality, here, I mean the unequal possession of cognitive abilities, epistemic resources such as concepts, and skills such as analytic thinking to assess the credibility of a piece of information. Citizens are unequally skilled and unequally trained in judging whether the source reporting news should be trusted, whether it is possible to trace credible evidence behind the news and, when implications are drawn, whether such implications follow from the reported facts. Epistemically disadvantaged citizens, thus, are more vulnerable to failing to identify fake news, believing it or judging it as possi-

bly true, and passing it on without even trying to assess its credibility. Note that such an epistemic disadvantage is not purely innate: epistemic abilities are highly dependent on training. Education contributes to training appropriate epistemic skills. Moreover, exposure to good-quality media also contributes to making news consumers more competent as knowers, while poor-quality media reduces consumers' epistemic abilities.

Epistemic inequality alone does not predict who believes or shares fake news. Yet empirical studies on the predictors of belief in and spread of fake news suggest that epistemic inequality does contribute to making some citizens more vulnerable. According to recent reviews (Baptista and Gradim 2022, 2; see also Baptista and Gradim 2020), vulnerability to fake news correlates with lower education, reduced analytic thinking and old age, which, I argue, can be proxies of epistemic disadvantage in discerning fake news on social media. People with a lower level of education may be epistemically disadvantaged to the extent that they have been less trained to discern reliable sources of information, to identify logical fallacies and inconsistencies in reasoning and so forth. Indeed, empirical literature also documented an association between belief in fake news and reduced analytic thinking (Bronstein et al. 2019). Analytic thinking promotes the ability to distinguish meaningful statements from statements constructed without concern for truth, but it is more effortful than intuitive thinking. Indeed, default responses are suggested by intuitive cognitive processes emerging autonomously from simple stimulus-response pairings (Bronstein et al. 2019). Hence, although some individuals may be more inclined to analytic thinking as a cognitive style, analytic thinking as an epistemic skill needs to be trained. Thus, citizens with lower levels of education, who have been less trained in analytic thinking, appear epistemically disadvantaged in discerning fake news and thus are more vulnerable to believing and sharing it.

Moreover, people who received a low level of education typically have a low occupation level and live on low wages. This may be relevant when it comes to their access to online information. Indeed, empirical data suggest that people with a low occupation level mainly consume news distributed via social media and are less likely to actively search for news on newspaper websites (Kalogeropoulos and Nielsen 2018). Those who, instead of directly accessing newspaper websites, receive their information primarily through the mediation of social media and instant messaging apps receive less detailed and lower-quality information, even when such distributed information originally comes from mainstream media. Indeed, newspapers and magazines typically reserve longer and more sophisticated articles on complex topics for paying

subscribers, while shorter, simpler, more frivolous and sensationalist articles are accessible for free on social media. Non-subscribers are thus fed with worse quality content and get used to increasingly simplified reasoning and language rather than to complexity. Free news articles from mainstream media may not dramatically differ from articles taken from alternative information sources and do not appropriately train the epistemic skills required to discern fake news from reliable information. Compared to genuine news articles, fake news articles are typically shorter, include fewer technical words and quotes, employ smaller words, use less punctuation, and show more lexical redundancy. Hence, fake news needs lower levels of education to be interpreted (Baptista and Gradim 2020, 9). When even the genuine news articles a person is used to read are free news articles, which are often simplistic and feature clickbait headlines, that person is disadvantaged in detecting oversimplification in a fake news story, invalid inferences, or a lack of reference to recognizable and reliable sources.

Reduced epistemic skills and limited digital literacy might also explain why vulnerability to believing fake news and the likelihood of diffusing it correlates with older age. During the 2016 US election, older adults' Twitter feeds contained the most fake news (Brashier and Schacter 2020). An experimental study in the US found that respondents aged over 65 shared seven times as many articles from fake news domains on Facebook as respondents between 18 and 29 years of age (Guess et al. 2019). Another experimental study on US and UK Facebook users confirmed that fake news had a much higher reach among older users. Very few recipients opened the link, but many dropped emotional comments below the post, seeming to take the headline at face value (Loos and Nijenhuis 2020). Research in psychology suggests that cognitive declines alone cannot explain older adults' engagement with fake news (Brashier and Schacter 2020). Given that older people typically show a more constant ideology and party identification than young people, partisan motivations may contribute to their higher rates of engagement and reposting of fake news. However, this may also depend on the fact that while older adults have gained increasing access to social media, [2] they may lack the level of digital media literacy necessary to reliably determine the trustworthiness of news encountered online (Guess et al. 2019; Brashier and Schacter 2020). As I noted in the introduction, social media have not been devised to inform but rather to entertain. Navigating and categorizing the contents one

[2] According to the Pew Research Center, in the US, 46% of people over the age of 60 used Facebook in 2018 (Loos and Nijenhuis 2020: 71).

comes across in social media requires specific epistemic skills, which older citizens may not have sufficiently been trained to develop.

Let me clarify that I am not suggesting that *all* those who are more susceptible to fake news do in fact believe them. I am simply highlighting that epistemic disadvantage makes some citizens more vulnerable than others to fall for fake news. Indeed, lower education, old age, and reduced analytic thinking (which can be taken as proxies of epistemic disadvantage) proved to correlate with higher rates of belief in fake news, though surely not all less educated citizens, older citizens, or citizens with a prevalent intuitive cognitive style believe fake news. Furthermore, I am not arguing that *only* less educated, older, or intuitive citizens believe fake news. The misplaced testimonial credibility given to news shared on social media, as highlighted by Rini (2017), proved stronger when fake news is passed on among people who share the same partisan affiliation. One is more likely to accept fake news if it is received from someone who shares similar worldviews, identifies with the same political party, or supports the same candidate. Furthermore, political partisanship proved to play a crucial role in strengthening biases and motivated reasoning (see Anderson 2021). Several studies suggest that motivating reasoning is particularly strong among right-wing voters and conservatives (Baptista and Gradim 2022). Yet empirical literature also reveals that right-wing and conservatives show lower levels of analytic or reflective thinking (Deppe et al. 2015) and a higher susceptibility to conspiracy theories (Douglas 2018). Thus, the evidence on partisanship as concurring to explain the unequal susceptibility to fake news among citizens does not exclude the role played by epistemic disadvantage. Epistemic disadvantage too might contribute to make part of the right-wing audience more inclined to accept to fake news as plausible.

Often, citizens who belong to socially disadvantaged social groups are also epistemically disadvantaged. Note, however, that individuals who count as privileged along ethnic or gender lines, such as white male Americans, may nonetheless be epistemically disadvantaged in knowledge acquisition and, specifically, in discerning fake news. Epistemic privilege itself is not unidimensional. Being given credibility as a source of testimonial knowledge is one of these dimensions, and members of socially dominant groups are typically privileged in this respect. For instance, white men in the US are typically epistemically privileged in their testimonial capacities compared to Black and female people, who are subjected to a systemic credibility deficit due to racist and sexist prejudice, suffering what Fricker (2007) famously defined as testimonial injustice. However, even white men can be epistemically disadvantaged in the distribution of epistemic goods, such as educational opportunities

in the development of their epistemic resources and in the acquisition of epistemic skills such as critical and analytic thinking.[3] In their case, believing fake news may well partly depend on prejudice in accommodating false information that fits previous beliefs and on willful ignorance in dismissing unfitting information (see Pohlhaus 2012). Yet, it also partly depends on factors such as low levels of education or low online media literacy and reduced access to the wider scientific community.

If some citizens are more susceptible to believing fake news and unwittingly misinform others due to their epistemic disadvantage, this means that epistemic inequality mediates the epistemic harm of fake news, making citizens unequally vulnerable to being harmed by fake news in their capacity as knowers. What I intend to highlight now, is the fact that the epistemic harm of believing fake news is particularly detrimental for those who were already epistemically disadvantaged, because it deepens their epistemic disadvantage. Consider the effect of fake news on someone who genuinely believes it to correspond to something that happened. This false information becomes part of their body of knowledge about the world. Thus, the fake news recipient comes to know less about the world. Often, a person accepts fake news as true or plausible because it confirms their worldviews, their stereotypes and prejudice. Thus, exposure to fake news makes them less informed about reality and even more biased towards future evidence because the incorporated false information now constitutes evidence to judge future information.

Furthermore, accepting an untrustworthy source of information as trustworthy undermines the ability to judge the credibility of future information. Suppose someone received a website link shared by a Facebook friend they trust, and the recipient believed that news to be true. Why would they be suspicious about an analogous article from a similar website, especially when shared by the same Facebook friend? Fake news recipients, then, become less competent in resisting fake news the more they attribute trust to unreliable sources of information. As we have seen, those who believe and pass on fake news may have already been disadvantaged in accessing trustworthy sources of information. They may have been prevalently fed with mediated information, rather than directly searching for news, and with free, low-quality news articles. Low-quality information may have eroded trust in mainstream media and made them more open to alternative sources of information, including those

[3] For a conception of epistemic injustice understood as an injustice in the distribution of epistemic goods such as information and education, see Coady 2010. On the concept of "formative epistemic injustice", see Nikolaidis 2021.

who nurture conspirative thinking. Falling for a piece of fake news feeds this unwarranted trust in alternative sources of information. Hence, fake news recipients are made more vulnerable to the next piece of misinformation they will be exposed to. They will not only know less about the world and have their stereotypes and prejudices rooted more deeply. They will also be less capable to discern false information and to discriminate among information sources. Shortly, they will be even more epistemically disadvantaged than before being exposed to fake news.

In sum, epistemic inequality mediates the impact of fake news on society. The epistemic harm it imposes on citizens as knowers has an asymmetric effect. Those who are disadvantaged as knowers before being exposed to fake news are more vulnerable to accepting it as true or plausibly true and to passing it on. Moreover, incorporating false information, invalid arguments, and unreliable sources in their body of knowledge makes them increasingly less epistemically competent as knowers, deepening their previous epistemic disadvantage. Theoretically speaking, taking pre-existing epistemic inequality into account allows us to have a better understanding of the epistemic harm to citizens as knowers. Normatively speaking, it is relevant for the debate on the responsibility of citizens who pass on fake news, misinforming their fellow citizens. Rather than too quickly placing blame on all of them for their complicity in spreading fake news, political philosophy should consider the multidimensional sources of epistemic inequality among citizens in accessing and processing information. This does not mean that citizens should not be held responsible for their behavior as misinformation consumers and misinformation spreaders. However, it entails paying attention to their asymmetric vulnerability to fake news and to the asymmetric detrimental impact that believing fake news has on those who already had fewer opportunities to develop the appropriate epistemic skills to discern this form of misinformation. Hence, when discussing how democracies should counter the spread of fake news, structural background inequalities in the development of appropriate epistemic skills should be taken into account, alongside individual responsibilities in consuming online content.

III. Oppressed minorities' unequal vulnerability to be targets of fake news

The previous section argued that the epistemic harm of fake news to citizens as knowers has an asymmetric impact on society because citizens are un-

equally exposed to this harm due to epistemic inequality. This section aims to show that some citizens, namely members of oppressed minorities, are also unequally exposed to being the object of fake news stories that reinforce their inferior status in society. When discussing the damages of fake news, we tend to focus on false stories about political candidates and other public figures. However, a large subset of fake news stories online revolves around members of minorities, such as immigrants, Muslims, Roma, and Jewish people. Such minorities are disproportionally vulnerable to being negatively represented in false stories spread on social media. A quantitative study on above 600 false stories circulated in the US, the UK, Germany, and Austria revealed that in Germany and Austria, fake news stories on immigrants even outnumbered those targeting political actors (Humprecht 2019). Moreover, immigration was a salient issue both during the 2016 US presidential elections campaign and during the UK Brexit referendum campaign. Indeed, anti-minority fake news is widely used in politics. Even fake news stories that directly target rival political candidates may in fact reinforce anti-minority sentiments as well. For instance, a fake news story reported by Trump's national security adviser, Michael Flynn, featured Democratic senators wanting to impose Sharia law in Florida (Borella and Rossinelli 2017). This was not only a piece of anti-Democrats fake news but also a contribution to Islamophobic sentiments.

Islamophobic and xenophobic sentiments appear strictly intertwined in online misinformation. Several pro-leave tweets pointed to a global conspiracy in which US President Obama and the German Chancellor Merkel were supporting a Muslim invasion of the West. Single fake news entries, such as those concerning Turkey joining the European Union (EU) and its free-movement area, fit in this larger conspirative frame. In the EU vs Disinfo database of anti-minority fake news, Muslims are the second main target after migrants (Szakács and Bognár 2021, 12). According to an analysis of this database, fake news tends to present migrants and Muslims primarily as a threat to the European culture. This is illustrated in false stories about European cities abandoning Christmas traditions or children forced to prey to Allah in schools (Szakács and Bognár 2021: 13). Secondly, fake news often portrays both migrants and Muslims as criminals, particularly as rapists (Szakács and Bognár 2021,13). This is illustrated in widely circulated fake news on the increase in crime in Germany, as well as on the increase of rapes in Sweden. [4]

[4] The latter is an example of false context misinformation rather than a case of a completely fabricated story: Swedish statistics do show an increase in reported rapes, but the uncontextualized sensationalist data reported in fake news omit that it depends on changes in rape definition

The Covid-19 pandemic seems to have lowered the overall salience of migration in the public debate in Europe, at least temporarily. Nevertheless, it boosted fake news that builds on the atavistic fear of strangers as disease carriers. Fake news about newly arrived immigrants escaping quarantine, violating lockdown restrictions, or even deliberately infecting police officers were reported in several countries, including Italy, Spain, Germany, and Belgium. Fake news about immigrants making the majority of intensive care patients circulated in Germany, too (Szakács and Bognár 2021, 16-17). Among citizens distrusting mainstream media and political elites, the sharp decline in media coverage about migration after the outbreak of the pandemic also elicited conspirative thinking. False stories on migrants entering the country under the cover of lockdown spread both in Germany and the Czech Republic (Szakács and Bognár 2021, 16). The pandemic also reinvigorated deeply rooted antisemitism and anti-Roma sentiments. Along with fake news on alleged Zionist plans for ethnic substitution and suppression of nation states, fake news spread on Jews being implicated in the creation and diffusion of the new virus. In several countries, Roma minorities were scapegoated on social media for the spread of Covid-19. In countries such as Slovakia and Bulgaria, where Roma minorities were subjected to additional movement restrictions, fake news resulted in direct harmful policy outcomes (Szakács and Bognár 2021, 15).

Fake news stories about immigrants, Muslims, Jews and Roma minorities are particularly relevant when assessing the damages of fake news on democratic societies. The content of anti-minority fake news has a specific harmful impact on the members of targeted minorities. Indeed, such instances of fake news reinforce the subordinate status of members of the targeted minorities, who are not treated as equals in society. Thus, they are distinctively wrongful to minority members. In the next section, I illustrate this claim by focusing on how anti-immigrant fake news contributes to the oppression of immigrant minorities. Furthermore, fake news reinforcing a pre-existing form of subordination is particularly damaging for the institutions of a society that aims to live up to its democratic values, given that moral equality among citizens is a fundamental democratic value and the abolition of status hierarchies is a key goal of democratic institutions.

and count in Sweden, and do not support a causal connection with the admission of immigrants and refugees (Juhász and Szicherle 2017).

IV. Amplifying oppression: the case of anti-immigrant fake news

As shown in the EU vs Disinfo database (Szakács and Bognár 2021, 12), anti-immigrant fake news accounts for most instances of anti-minority fake news in Europe. Anti-immigrant fake news is distinctively harmful and wrongful because it worsens the disadvantaged position of immigrant minorities in Western liberal democracies. When talking of immigrant minorities, here, I do not only refer to foreign residents. I also refer to naturalized citizens and citizens whose parents or grandparents immigrated who continue to be identified as "immigrants" and to occupy a subordinate position in society because of their negatively stereotyped ethnic origins or religious belonging. Even when it targets potential immigrants, anti-immigrant fake news is harmful to members of immigrant minorities who are already members of society.

The social disadvantage affecting immigrant minorities is not reducible to inequalities in income and wealth, although members of immigrant minorities are often more exposed to material deprivation. The social inequality at stake is broader: it is a status inequality, a disadvantage in their standing as members of society. Status inequality does not merely reflect the distribution of material goods. It depends on how the minority group is commonly perceived and publicly represented, on how its members are treated by institutions in laws and policies, but also by their fellow citizens in everyday practices. For newly arrived immigrants and asylum seekers, social inequalities may depend on their legal status: indeed, undocumented migrants and legal residents with temporary visas may not be granted the same rights as citizens. Yet, for foreign-born citizens and for second-generation or third-generation immigrants, social inequality may persist despite access to equal rights.

To see the forms of inequality that members of immigrant minorities experience and why they amount to social injustice, considering the distribution of resources and opportunities is not sufficient. It is necessary to take institutionalized norms and social relations into account. Among the accounts of non-distributive forms of social injustice, Iris Marion Young's (1990) theory of oppression as a multifaced phenomenon is a prominent example. Moreover, it offers a sophisticated frame to make sense of how anti-immigrant fake news reinforces the disadvantaged position of immigrant minorities. Drawing on Young's account, I argue that members of immigrant minorities in Western liberal democracies often experience marginalization, cultural imperialism and violence, which are forms of oppression. This allows us to capture the impact of anti-immigrant fake news on targeted immigrant minorities. Indeed, anti-

immigrant fake news contributes to their oppression, further eroding their already unequal standing in society.

A striking way in which anti-immigrant fake news denies equal status to members of targeted minorities consists of an expressive wrong. Think of fake news stories that depict people with an immigrant background as criminals or Muslims as unable to respect basic liberal democratic principles. Examples of anti-immigrant fake news of this kind contribute to amplifying the negative stereotyping and essentialization that members of immigrant minorities already suffer. Members of denigrated immigrant minorities are misrecognized as individuals, reduced to an undifferentiated collective entity showing invariable negative traits. Fake news that associates immigrants with threatening behaviors and dangerous beliefs makes them even more visible in the media and among the public as members of a negatively stereotyped minority. Paradoxically, even debunking fake news may reiterate their negative stereotypical representation. At the same time, members of immigrant minorities, as individuals with their own views and experiences, are silenced.

This can be described, in I. M. Young's terms, as "cultural imperialism". According to Young (1990, 59), those living under cultural imperialism "undergo a paradoxical oppression, in that they are both marked out by stereotypes and at the same time rendered invisible". Anti-immigrant fake news amplifies this "face of oppression" when it reinforces negative stereotypes of immigrants as inferior and deviant and foster prejudice towards them. Moreover, those who undergo cultural imperialism cannot but perceive themselves from the perspective of those who negatively stereotype and demean them. They "find themselves defined from the outside […]. Consequently, the dominant culture's stereotyped and inferiorized images of the group must be internalized by group members at least to the extent that they are forced to react to behavior of others influenced by those images" (Young 1990, 59-60). Due to the anonymity and pervasiveness of anti-immigrant fake news, resisting internalization and rebutting demeaning stereotypes has become increasingly difficult for the members of targeted immigrant minorities.

Empirical research has shown that negative stereotypes and prejudice towards immigrants, particularly those belonging to certain ethnic and religious minorities, have an impact on behavior. Fake news that nurtures negative stereotypes and prejudice towards immigrants is wrongful in itself, but it also contributes to fueling other forms of oppression, namely marginalization and violence. At the beginning of the 1990s, Young identified a "growing underclass of people permanently confined to lives of social marginality, most of whom are racially marked", including "Blacks, East Indians, Eastern Europe-

ans, or North Africans in Europe". Marginalization remains a useful concept to describe the interplay between forms of social inequality, such as employment discrimination, and de-facto social and spatial segregation, that many immigrants face today. According to Young (1990, 53), marginalization can be the most dangerous form of oppression, whereby "a category of people is expelled from useful participation in social life and thus potentially subjected to severe material deprivation". Not all individuals, even among the most despised ethnic and religious minorities, suffer marginalization to the same extent. Yet, it can still be argued that, as a social group, they are systematically "exposed to deprivation of cultural, practical, and institutionalized conditions for exercising capacities in a context of recognition and interaction" (Young 1990, 55).

Several studies have shown that immigrants face discrimination in employment and housing. This emerges consistently from field experiments based on correspondence audits, in which researchers send written applications for employment or housing in response to real job or housing advertisements, manipulating information about the correspondents. Research both in Europe and North America demonstrated evidence of employment discrimination toward immigrants applying for jobs. Holding human capital constant, stronger discrimination emerges against African, North African, Middle Eastern and Muslim applicants (Esses 2021, 512-514). Correspondence audits have also been used to study housing discrimination, focusing on responses to rental advertising. This literature, albeit less extensive than the literature on employment discrimination, shows that there is prevalent discrimination against Middle Eastern, North African, and Arab Muslims in Europe, as well as against Hispanics in the US (Esses 2021, 514-515). As a result of discrimination in employment, immigrants belonging to these ethnic and religious minorities find themselves systematically more unemployed or underemployed than the rest of the population (Esses 2021, 519). Moreover, housing discrimination contributes to the creation of neighborhoods with a high concentration of immigrants, which are often those with fewer services and resources. Hence, immigrants have poorer access to good jobs and high-quality schools, they need to commute longer to work and have more difficult access to medical care (Esses 2021, 520).

The widespread circulation of fake news portraying members of minorities as deviant and inferior normalizes and provides justification for discriminatory behavior and social exclusion. If so many pieces of news associate members of targeted minorities with certain threats, even those citizens who were not strongly prejudiced against the targeted minority may suspect that there might

be something true and behave accordingly. Despite establishing causation between anti-immigrant fake news and discriminatory behavior can be difficult, a recent study by the Spanish Ministry of Equality linked the spread of fake news and online hate speech to a rise in anti-immigrant discrimination in housing and education (Szakács and Bognár 2021, 23). Unemployment and underemployment due to discrimination in job seeking and spatial confinement in underserviced neighborhoods due to housing discrimination, in turn, contribute to material deprivation and limited social relationships with the rest of society.

Fake news, thus, contributes to undermining the status of members of negatively connotated immigrant minorities within society by amplifying their marginalization. What is more, anti-immigrant fake news fosters another serious form of oppression, which is systemic violence. According to Young (1990, 62), what makes violence "a phenomenon of social injustice, and not merely an individual moral wrong, is its systemic character, its existence as a social practice. Violence is systemic because it is directed at members of a group simply because they are members of that group." Members of negatively marked groups, Young notes, have a "daily knowledge" they are "liable to violation, solely on account of their group identity." The risk of being physically or verbally attacked may well be independent of their behavior, given that this systemic violence is motivated by fear and hatred towards the group they belong to. Young argues that the wrong of violence does not only apply in the moment one is individually attacked: the mere fact of "living under such a threat of attack on oneself or family or friends deprives the oppressed of freedom and dignity" (Young 1990, 62). Fake news fuels both online and offline hate crimes. Not only it incites online hate speech, but also verbal and physical violence in in-person interactions. Although a direct causal connection between the circulation of a piece of fake news and offline hate crimes may be hard to establish, recent empirical studies support the claim that an increase in online hate speech on social media is associated with an increase in violence towards the targeted minority. In their study on hate crimes in London, Williams et al. (2020) established a "temporal and spatial association between online hate speech targeting race and religion and offline racially and religiously aggravated crimes". In the German context, Müller and Schwarz (2021) found a link between antirefugee sentiment on Facebook and hate crimes against refugees.

In sum, anti-immigrant fake news exacerbates marginalization, cultural imperialism and violence and thus contributes to reinforcing the overall oppression of the members of targeted immigrant minorities. Thus, anti-

immigrant fake news has an asymmetric impact on citizens: it is distinctively harmful to those citizens who belong to targeted immigrant minorities. Furthermore, the fact that anti-immigrant fake news worsens the inferior social status of members of immigrant minorities is particularly worrisome based on a democratic conception of political justice. Indeed, fake news stories that reinforce the oppression of members of immigrant minorities contribute to hindering their capacity to stand as equals and participate in the democratic political process, strengthen unjust social hierarchies among citizens and further erode the democratic value of status equality.

Conclusion

In this paper, I argued that part of what makes fake news problematic for liberal democratic societies is the role fake news plays in amplifying pre-existing social inequalities and injustices. It has already been shown that fake news is harmful to recipients in their capacity as knowers. I argued that this epistemic harm does not have a uniform impact on democratic societies, given the epistemic inequality among citizens. Those who are epistemically disadvantaged are more susceptible to believing fake news. Believing fake news, in turn, deepens this epistemic disadvantage. Secondly, I argued that citizens are unequally targeted by fake news and that targeted groups are distinctively harmed. Indeed, a significant subset of fake news consists of misinformation about minorities, such as immigrants, Muslims, Jews, and Roma minorities. Such anti-minority fake news is distinctively harmful to members of targeted minorities. I focused on the case of anti-immigrant fake news to show how anti-minority fake news amplifies the status inequality of the targeted minority within society. Drawing on I. M. Young's account of oppression, I claimed that anti-immigrant fake news exacerbates multiple dimensions of oppression, namely marginalization, cultural imperialism, and violence against members of targeted immigrant minorities. Finally, I noted that anti-minority fake news is not only notable for its unequal damaging impact, falling disproportionally on members of targeted minorities, but also for its particularly corrosive effect on democratic values. Indeed, the spread of fake news that reinforces the pre-existing oppression of a minority further erodes the democratic value of citizens' status equality.

Works Cited

Anderson, Elizabeth. 2021. "Epistemic Bubbles and Authoritarian Politics." In *Political Epistemology*, edited by Elizabeth Edenberg and Michael Hannon. Oxford: Oxford University Press.

Baptista, João Pedro and Gradim, Anabela. 2020. "Understanding Fake News Consumption: A Review." *Social Sciences* 9 (10):185.

Baptista, João Pedro and Gradim, Anabela. 2022. "Who Believes in Fake News? Identification of Political (A)Symmetries." *Social Sciences* 11 (10): 460.

Borella, Carlo Alessandro. 2017. "Fake News, Immigration, and Opinion Polarization." *SocioEconomic Challenges* 1 (4): 59-72.

Brashier, Nadia M and Schacter, Daniel L. 2020. "Aging in an Era of Fake News." *Current Directions in Psychological Science* 29 (3): 316-323.

Coady, David. 2010. "Two concepts of epistemic injustice." *Episteme* 7 (2): 101-113.

Bronstein, Michael V., Pennycook, Gordon, Bear, Adam, Rand, David G. and Cannon, Tyrone D. 2019. "Belief in Fake News is Associated with Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking." *Journal of Applied Research in Memory and Cognition* 8 (1): 108-117.

Brown, Étienne. 2021. "Regulating the Spread of Online Misinformation." In *The Routledge Handbook of Political Epistemology*, edited by Michael Hannon and Jeroen de Ridder. Abingdon: Routledge.

Chambers, Simone. 2021. "Truth, Deliberative Democracy, and the Virtues of Accuracy: Is Fake News Destroying the Public Sphere?" *Political Studies* 69 (1): 147-163.

Croce, Michel and Piazza, Tommaso. 2021. "Misinformation and Intentional Deception." In *Virtues, Democracy and Online Media: Ethics and Epistemic Issues*, edited by Nancy E. Snow and Maria Silvia Vaccarezza. London: Routledge.

Deppe, Kristen D., Gonzalez, Frank J., Neiman, Jayme L., Jacobs, Carly, Pahlke, Jackson, Smith, Kevin B. and Hibbing, John R. 2015. "Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology." *Judgment and Decision Making* 10: 314-31.

Douglas, Christopher. 2018. "Religion and Fake News: Faith-Based Alternative Information Ecosystems in the US and Europe." *The Review of Faith & International Affairs* 16 (1): 61-73.

Esses, Victoria. M. 2021. "Prejudice and Discrimination Toward Immigrants." *Annual Review of Psychology* 72 (1): 503-531.

Fricker, Miranda. 2007. *Epistemic injustice: Power and the ethics of knowing*. New York: Oxford University Press.

Galeotti, Anna Elisabetta and Meini, Cristina. 2022. "Scientific Misinformation and Fake News: A Blurred Boundary." *Social Epistemology* 36 (6): 703-718.

Guess, Andrew, Nagler, Jonathan and Tucker, Joshua. 2019. "Less than you think:

Prevalence and predictors of fake news dissemination on Facebook." *Science Advances* 5: eaau4586.
Hannon, Michael. 2021. "Disagreement or Badmouthing? The Role of Expressive Discourse in Politics." In *Political Epistemology*, edited by Elizabeth Edenberg and Michael Hannon. Oxford: Oxford University Press.
Humprecht, Edda. 2019. "Where 'fake news' flourishes: a comparison across four Western democracies, Information." *Communication & Society* 22 (13): 1973-1988.
Jaster, Romy and Lanius, David. 2018. "What is Fake News?" *Versus* 127: 207-227.
Juhász, Attila and Szicherle, Patrick. 2017. "The Political Effects of Migration-Related Fake News, Disinformation and Conspiracy Theories in Europe." Friedrich Ebert Stifüng; Political Capital.
Kalogeropoulos, Antoni and Nielsen, Rasmus. K. 2018. "Social Inequalities in News Consumption." Factsheet report. Reuters Institute for the Study of Journalism. Available at: https://reutersinstitute.politics.ox.ac.uk.
Levy, Niel. 2017. "The Bad News About Fake News." *Social Epistemology Review and Reply Collective* 6: 20-36.
Lewandowsky, Stephan, Ecker, Ullrich K.H., Seifert, Colleen M., Schwarz, Norbert and Cook, John. 2012. "Misinformation and its Correction: Continued Influence and successful Debiasing." *Psychological Science in the Public Interest* 13: 106-131.
Loos, Eugene and Nijenhuis, Jordy. 2020. "Consuming Fake News: A Matter of Age? The perception of political fake news stories in Facebook ads." In *Human Aspects of IT for the Aged Population*, edited by Quin Gao and Jia Zhou. Cham: Springer.
Müller, Karsten and Schwarz, Carlo. 2021. "Fanning the Flames of Hate: Social Media and Hate Crime." *Journal of the European Economic Association* 19 (4): 2131-2167.
Nikolaidis, Alexandros C. 2021. "A Third Conception of Epistemic Injustice." *Studies in Philosophy and Education* 40: 381-398.
Pohlhaus, Gaile Jr. 2012. "Relational Knowing and Epistemic Injustice: Toward a Theory of Willful Hermeneutical Ignorance." *Hypatia* 27 (4): 715-735.
Regliz, Merten. 2022. "Fake News and Democracy." *Journal of Ethics and Social Philosophy* 22 (2): 162-187.
Rini, Regina. 2017. "Fake News and Partisan Epistemology." *Kennedy Institute of Ethics Journal* 27 (2): E-43-E-64.
Rini, Regina. 2021. "Weaponized Skepticism: An Analysis of Social Media Deception as Applied Political Epistemology." In *Political Epistemology*, edited by Elizabeth Edenberg and Michael Hannon. Oxford: Oxford University Press.
Szakács, Judith and Bognár, Éva. 2021. *The impact of disinformation campaigns about migrants and minority groups in the EU*. Available at: https://www.europarl.europa.eu.

Williams, Matthew L., Burnap, Pete, Javed, Amir, Liu, Han and Ozalp, Sefa. 2020. "Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime." *The British Journal of Criminology* 60 (1): 93-117.
Young, Iris Marion. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University Press.

DOI 10.17473/2240-7987-2024-2-12

Uncertainty and Fake News: An Experimental Study on the Strategic Use of Fake News in Belief Formation*

Irene Maria Buso**, Margherita Benzi***, Marco Novarese****, Giacomo Sillari*****

Abstract

This paper investigates the strategic role of low-informative signals – conceptualized as fake news – in shaping belief formation and influencing decision-making under uncertainty. Building on the experimental frameworks of Exley (2016) and Garcia et al. (2020), we introduce a novel design that mimics real-world misinformation through signals that are true but partial, fostering biased beliefs. Specifically, we test whether such signals influence individuals to favor self-serving options, particularly when the misleading information aligns with their self-interest. Our results show strong evidence of excuse-driven behavior under risk-for-self, which is exacerbated by low-informative signals. Notably, the bias induced by weak signals diminishes when outcomes primarily benefit others, highlighting the strategic alignment of belief distortion with self-interest. These findings underscore the broader philosophical and empirical importance of understanding how partial information influences motivated reasoning and decision-making, contributing to the literature on misinformation, motivated beliefs, and behavioral economics.

* We thank CESARE Lab for the use of the laboratory facilities and the audience of the workshop The Democratic Containment of Fake News and Bad Beliefs held in LUISS in October 2023 for valuable comments and insights. Research realized with the contribution of the Italian Ministry of University and Research as part of the PRIN project 2017, Deceit and Self-Deception. How We Should Address Fake News and Other Cognitive Failures of the Democratic Public, 2017S4PPM4_004.

** ORCID: 0009-0009-2828-1602.

*** ORCID: 0000-0003-4934-6494.

**** ORCID: 0000-0001-5622-5549.

***** ORCID: 0000-0002-3243-1206.

Summary: Introduction. – I. Experimental design and research questions. – II. Results. – Conclusions and directions for future research. – Works Cited. – Appendix.

Introduction

Our study explores whether low-informative signals—akin to fake news—strategically shape beliefs and behavior, particularly when they support self-serving choices. The present experimental analysis aims to shed light on the relationship between the formation of motivated beliefs and fake news, defined here as partial, low-informative signals that can be misleading. This mechanism reflects a broader dynamic of misinformation, where selective or ambiguous evidence fosters doubt and skews decision-making. Philosophers Cailin O'Connor and James Owen Weatherall (2019) argue in *The Misinformation Age* that misinformation campaigns strategically exploit uncertainty to manipulate public belief. Real-world cases, such as the tobacco industry downplaying the causal link between smoking and cancer (Oreskes & Conway, 2010; Michaels, 2008), or climate change denial tactics that amplify minor uncertainties to obscure scientific consensus (Supran & Oreskes, 2017; Dunlap & McCright, 2011), demonstrate how sowing doubt can delay meaningful action. Similarly, misinformation surrounding vaccines and autism, which gained traction through the fraudulent Wakefield study, highlights how selective and misleading evidence can spread widely despite overwhelming contradictory data (Godlee et al., 2011). These real-world cases demonstrate how partial, ambiguous or low-informative signals, while technically true, can distort belief formation in ways that align with strategic self-interest, opposing public interest.

By experimentally simulating this phenomenon, we provide an empirical investigation of how partial or ambiguous signals—like those found in misinformation campaigns—can shape belief formation and justify self-serving choices. Our study, therefore, contributes to a growing philosophical and empirical literature on the strategic role of misinformation in fostering biased reasoning and motivated beliefs. Specifically, we test whether fake news (uninformative signals) about the consequences of a choice influences decision-making toward selfish rather than prosocial options, particularly when the misinformation aligns with self-interest. This experimental work is of philosophical significance as it helps to identify and isolate the conditions under which uncertainty and low-informative signals are exploited, mirroring the mechanisms underlying misinformation in broader societal contexts.

Previous experimental evidence has suggested that uncertainty on the consequences of a choice on the welfare of other individuals (e.g., Dana et al., 2007; Exley, 2016; Garcia et al., 2020) or on the environment (e.g., Momsen and Ohndorf, 2022), as well as uncertainty on the prevalent social norm (e.g., Bicchieri et al., 2023), creates a moral wiggle room that is exploited to undertake selfish behavior. Belief manipulation has been identified as a source of self-serving behavior (e.g., Bicchieri et al., 2023; Hasley and Weber, 2010). Our experiment belongs to this strand of literature, as it aims to test whether fake news about the consequences of a choice on the welfare of another individual influence belief formation and decision-making towards selfish rather than prosocial options. Specifically, we examine if fake news induces biased beliefs, particularly when the misinformation aligns with self-interested choices, thereby promoting motivated beliefs.

We build on the experimental setting of Exley (2016) and Garcia et al. (2020) to detect the strategic use of uncertainty to take more selfish choices. Four scenarios are implemented using multiple-choice lists, where subjects choose between a safe option and a risky option. These four scenarios differ in whether the beneficiary of the safe option and the risky option is the same subject making the choice or another recipient, such as another individual in the experiment or a charity. We add to this setting a signal on the payoff—either positive or null—that the lottery would determine. This signal could be informative to different degrees. Similarly to Poinas et al. (2012), subjects receive a true but imperfect signal on the urn composition, and the informativeness of the signal is exogenously manipulated. Subjects are told that one of two urns is randomly selected (each with a probability of 0.5) for each price list to determine the result of the lotteries. The white urn is composed of 1000 white balls, while the black urn contains 999 white balls and 1 black ball. The lotteries in a price list pay the positive outcome if the white urn is associated with that price list. The computer inspects the urn, checking the color of n balls with replacement, and provides a signal regarding the presence of the black ball. Depending on *n*, the signal on the absence of the black ball varies.

We suggest defining fake news as partial information that could be misleading, akin to political propaganda or selective economic reporting. This strategic use of low-informative signals mirrors how misinformation skews beliefs in real-world contexts. Experimental data show that self-serving behavior is present and is increased by the informative signal.

In Section II, we present the experimental design and research questions in detail. Section III provides the results, and Section IV concludes with directions for future research.

I. Experimental design and research questions

In this experimental analysis, we build upon the design of Exley (2016) to elicit excuse-driven responses to risk, which was then replicated by Garcia et al. (2020) and extended to ambiguity. To test whether uncertainty is used as an excuse not to give, participants make choices in four different scenarios. These scenarios are implemented through multiple price lists where subjects choose between a safe option and a risky option. We emphasize how low-informative signals in our experimental setup mirror real-world fake news scenarios, such as ambiguous or selective news items that skew public perception. For instance, the low-informative signal in our study is analogous to partial news reports that selectively present facts to influence audience beliefs. This is the conceptual connection between laboratory conditions and real-world misinformation.

As illustrated in Figure 1, the four scenarios differ in terms of who benefits from the safe option and the risky option:

Scenario 1: The beneficiary of both the safe and risky options is the same subject making the choice.

Scenario 2: The beneficiary of both options is another subject for whom the decision-maker is choosing.

Scenario 3: The decision-maker is the beneficiary of the risky option, while another subject benefits from the safe option.

Scenario 4: The decision-maker benefits from the safe option, while another subject benefits from the risky option.

For each scenario, it is provided a price list with 21 choices, and choice number 10 is the one used as an example in Figure 1.

Figure 1- The four scenarios

Scenario 1

Decision	Option A	Option B	Your Choice
10	You: lottery; other: 0$	You: 4.5$; other: 0$	Option A Option B

Scenario 2

Decision	Option A	Option B	Your Choice
10	You: 0$; other: lottery	You:0$; other: 8.1$	Option A Option B

Scenario 3

Decision	Option A	Option B	Your Choice
10	You: lottery$; other: 0$	You:0$; other: 8.1$	Option A Option B

Scenario 4

Decision	Option A	Option B	Your Choice
10	You: 0$; other: lottery	You:4.5$; other: 0$	Option A Option B

In each scenario, participants face a price list with 21 choices, varying in the value of the safe amounts and positive lottery outcomes depending on the scenario. The detailed technical descriptions of urn compositions, sampling procedures, and calibration tasks have been moved to the Appendix.

The novel aspect of our design involves the introduction of a signal about the realization of the lotteries, that is whether the lotteries in the price list yield the positive or null payoff. The signal, which can be more or less informative, is derived from observing a number of draws (n) from an urn. The draws can be relative to one of two urns—the black urn (associated to the null payoff) and the white urn (associated with the positive payoff). The urns are associated with equal probability (P(W)=P(B)=.5) to the price list. The white urn contains 1000 white balls, while the black urn contains 999 white balls and 1 black ball. The signal consists in observing *n* random draw with replacement from the urn actually assigned to the price list, allowing subjects to form a

posterior belief based on Bayes' rule regarding the likelihood of the urn being white. For example, if $n = 1$ and a white ball is observed, the signal is only weakly informative, suggesting the urn is slightly more likely to be white [1]. However, as n increases, the informativeness of the signal also increases. If a black ball is observed, the urn is conclusively determined to be the black urn.

The effect of the signal is assessed with a within-subject design. After a calibration task, participants are presented with each of the 4 scenarios, each one with no signal, a low informative signal (n=1,2,3) or a high informative signal (n=551,552,553). Overall participants are presented with 12 price lists. In particular, the signal consists of the following statement: "n balls have been randomly drawn, and no black ball has been found".

The 12 price list are presented in random order, and the order varies randomly between subjects. Our main hypotheses are:

1. *Excuse-driven behavior under risk-for-self:* We expect stronger excuse-driven behavior under risk-for-self when subjects are the beneficiaries of the risky option. Specifically, subjects are expected to be less risk-averse when the safe amount is assigned to another participant (Scenario 3) compared to themselves (Scenario 1). This hypothesis aims to replicate the findings of Garcia et al. (2020) and Exley (2016).
2. *Excuse-driven behavior under risk-for-other:* We expect weak evidence of excuse-driven behavior when the beneficiary is another subject. In Scenario 4, where the safe amount benefits the decision-maker and the risky option benefits another subject, subjects are expected to switch to the safe option earlier. This hypothesis also aims to replicate the findings of Garcia et al. (2020) and Exley (2016).

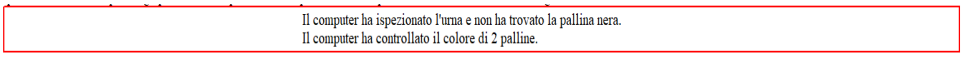
The novelty of our design concerns the introduction of the signal, what should reduce the uncertainty of the choice. Still, the imperfection of the signal, in particular, the low informative signal, could reinforce excuse driven behavior. Indeed, the low informative signal informing the subject that 1 (2 or 3) ball(s) have been drawn and there is not the black ball could foster the belief that the urn is the white one. Figure 2 present an example of the low informative signal (2a) and highly informative signal (2b).

[1] If for example the subjects observe n=1 and the random drawn is a white ball, the posterior probability that the urn is white P(W | n=1 white ball) is almost identical to the prior P(W): indeed, the signal is very weekly informative:

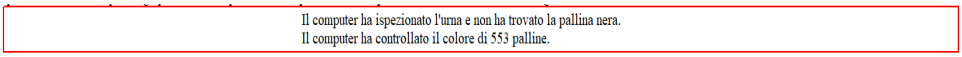
Pn=1 white ball=PW*PWPW*PW+PB*PB=0.5002513

where P(n=1 white ball | W)=1 and P(n=1 white ball | B)=0.999.

Figure 2 - The signal



2a- The low informative signal



2b- The highly informative signal

The hypothesis that we want to test regarding the low informative signal are the following:

3. *Impact of low-informative signals*: We aim to test whether weakly in formative signals affect decision-making in a self-serving manner, by fostering an overestimation of the likelihood of favorable outcomes. This effect is akin to the strategic use of ambiguous information in real-world fake news to reinforce self-serving narratives. Comparing risk taking for self (Scenario 1) without signal and with low informative signal, we expect to observe greater risk taking when the low informative signal of a white ball drawn is given.

Evidence of greater risk taking for self with the low informative signal would imply that subjects overestimate this signal respect to the Bayesian updating benchmark when the payoff consequences of the choice is for themselves, what means that there is no excuse driven motivation in such overestimation. However, our main interest is whether the low informative signal is used in a self-serving manner:

4. *Excuse-driven behavior under low-informative signals*: we expect that the low informative signal increases excuse driven behavior under risk for self. In this case the low informative signal reinforces this excuse driven behavior as it leads to overweight the probability of the positive outcome realization in a self-serving manner.

On the other hand,

5. We do not expect to observe a reinforcement effect of excuse-driven behavior (if any) when trading off payoffs for self and the recipient when the lottery benefits the recipient. Specifically, low-informative signals are not expected to increase excuse-driven behavior when the benefit is primarily for another individual, as opposed to oneself.

Hence, beyond replicating previous evidence (hypotheses 1 and 2) we aim to test whether weakly informative signals affect choices (hypothesis 3), and,

if so, whether they affect choices in a self-serving manner, that is trusting questionable aligned evidence (hypotheses 4 and 5).

In relation to fake news, the low-informative signals in our experimental design simulate the impact of partial or misleading information, similar to how selective news items can manipulate public beliefs. By providing subjects with weakly informative signals, we assess whether they use these signals to justify self-serving decisions, thereby demonstrating the influence of fake news in shaping behavior.

More precisely, depending on *n*, the signal on the information on the absence of the black ball is more or less informative. While we do not directly measure beliefs, we evaluate the impact of this information on subjects' choices [2]. Information structures have many features that may be used to distort beliefs. For example, Enke and Zimmermann (2019) show that a typical features of information structures of news media such as correlation between information determine distorted beliefs as subjects tend to neglect correlations between signals even in the simple environment of the lab. Another features of the information structures that could distort beliefs is the presentation to the public of true but partial information. If a partial information, i.e. a weakly informative piece of information as in n=1,2,3, is able to change behaviors these partial information can be used to create fake news. We decide to study then the impact of weakly informative information given the recent evidence that weakly informative signals change beliefs (Augenblick et al., 2021).

Finally, at the beginning of the experiment every subject is paired with another participant, and each participant of the pair has either a probability of 90% or 10% to be the dictator in a dictator game [3], and other group member has complementary probability. According to previous evidence of fairness in behavior with uncertainty in roles (e.g., Mesa-Vázquez, E., Rodriguez-Lara, 2021), we expect (research question 6) greater excuse driven behavior when the probability of being a dictator is 90% than when is 10%.

The experiment was conducted in September 2023 at the CESARE Lab at LUISS University. The sessions were programmed in z-Tree (Fischbacher, 2007); and participants were recruited using ORSEE (Greiner, 2015). A total

[2] As our main focus is on the use of weakly informative signals ("fake news") in a self-deceptive mechanism, the explicit reference to belief formation process may weaken this effect as it could be run unconsciously.

[3] The dictator game is a paradigmatic lab experiment in which a player (the "dictator") is given a sum of money to allocate between herself and a second player (the "recipient"). The recipient has no saying and receives the money allocated to her by the dictator.

of 90 subjects were recruited, with 72 completing the experiment [4]. They received a show-up fee of 5 euros, along with additional payments based on the outcomes of two randomly selected price lists.

II. Results

In this section, we present the results of the experimental analysis. We begin by analyzing the behavior of participants with a 90% probability of being assigned the role of dictator, followed by those with a 10% probability.

Behavior with 90% Probability of Being Dictator: Table 1 presents the average number of "Option A" choices (risky choices) made in Scenarios 1 and 3 by participants with a 90% probability of being the dictator. The observed difference of -6.33 between the number of risky choices in Scenario 1 and Scenario 3 replicates previous evidence of excuse-driven behavior under risk-for-self (Hypothesis 1). Moreover, the excuse-driven behavior under risk-for-self significantly increases when subjects receive a low-informative signal, as evidenced by the difference of -10.11 (Hypothesis 3). This indicates that the presence of a weak signal encourages participants to make more self-serving decisions by fostering overconfidence in the lottery outcome, similar to how fake news can manipulate beliefs.

Table 1

Feedback	Scenario 1	Scenario 3	
No	10.33 (0.44)	16.66 (1.49)	-6.33
Low	8.22 (1.01)	18.33 (0.86)	-10.11
This table present the average number of "Option A", i.e., risky choices, in Scenario 1 and 3 by players with 90% probability of being the dictator.			

Behavior with 10% Probability of Being Dictator: Table 2 shows the average number of "Option A" choices made in Scenarios 1 and 3 by participants with a 10% probability of being the dictator. Similar to the findings for participants with a 90% probability, there is evidence of excuse-driven behavior under risk-for-self; however, the effect is less pronounced, as indicated by the

[4] 12 subjects from the first session of 24 subjects dropped from experiment due to technical problems with the software, and their data are not considered in the analysis.

difference of -1.29 (Hypothesis 6). This aligns with our expectations that lower certainty in the role reduces the justification for self-serving behavior.

Table 2

Feedback	Scenario 1	Scenario 3	
No	11.42 (1.79)	12.71 (2.62)	-1.29
Low	11.71 (2.89)	14.7 (2.38)	-2.99
This table present the average number of "Option A", i.e., risky choices, in Scenario 1 and 3 by players with 10% probability of being the dictator			

Excuse-Driven Behavior under Risk-for-Others: For participants with a 90% probability of being the dictator, Table 3 shows weak evidence of excuse-driven behavior under risk-for-others, which slightly increases with the low-informative signal. The low-informative feedback leads to an increase in choosing the risky option for other players, with a more pronounced effect in Scenario 2 compared to Scenario 4. This supports Hypothesis 4, indicating that low-informative signals can reinforce motivated beliefs in situations where decision-makers act on behalf of others.

Table 3

Feedback	Scenario 2	Scenario 4	
No	8.33 (1.09)	5.88 (1.41)	2.45
Low	10.22 (1.89)	6.77 (2.19)	3.45
This table present the average number of "Option A", i.e., risky choices, in Scenario 2 and 4 by players with 90% probability of being the dictator			

Excuse-Driven Behavior with Low Probability: Table 4 presents data for participants with a 10% probability of being the dictator. There is stronger evidence of excuse-driven behavior under risk-for-others without feedback; however, this behavior does not emerge with low-informative feedback. This difference may be attributed to the randomness in decision-making by participants with a lower likelihood of becoming the dictator, reducing the robustness of the results when weak signals are presented. Future research should seek to validate these findings with larger samples to assess the stability of excuse-driven behavior under varying probabilities.

Table 4

Feedback	Scenario 2	Scenario 4	
No	12.42 (1.52)	6.85 (1.95)	5.57
Low	7 (2.7)	7 (1.51)	0
This table present the average number of "Option A", i.e., risky choices, in Scenario 2 and 4 by players with 10% probability of being the dictator			

The results presented in this section provide strong support for the hypotheses outlined in Section II. Specifically, we replicate previous findings of excuse-driven behavior under risk-for-self (Hypotheses 1 and 2), while extending our understanding to show that low-informative signals, akin to fake news, have a substantial effect in promoting motivated beliefs (Hypotheses 3 and 4). Furthermore, the role certainty (90% vs. 10% probability of being the dictator) significantly influences the magnitude of excuse-driven behavior (Hypothesis 6). These findings contribute to our understanding of how partial or misleading information, like fake news, can strategically shape belief formation and decision-making.

Conclusions and directions for future research

The experimental design presented aims to explore the role of weakly informative signals, which approximate fake news, in forming beliefs and affecting pro-social choices. Specifically, we aimed to shed light on the possible reinforcement of excuse-driven behavior documented in previous literature (Exley, 2016; Garcia et al., 2020). The data analysis shows that the low informative signal indeed reinforces excuse-driven behavior, particularly under risk-for-self and for players with a high probability of being the actual dictators. Several limitations of the present experimental design emerged from the data analysis. First, only one-third of the data collected does not exhibit multiple switches in the price list. Additionally, a few subjects continued to choose the lottery even when the black ball had been randomly drawn from the urn. These two elements suggest that some subjects may not have fully understood the rules of the experiment and its incentives. Although the design and instructions closely resemble the settings of Exley (2016) and Garcia et al. (2020), several modifications introduced in our study might have rendered the experiment more complex, including the feedback structure and the introduction of role uncertainty for being the dictator, as the donation does not go to a charity.

Despite these limitations, the study contributes valuable insights into the strategic use of weakly informative signals and their impact on self-serving behavior. It highlights the parallels between controlled laboratory settings and real-world scenarios where misinformation influences decision-making. Future research should focus on refining the experimental setup to address the identified limitations. Potential avenues for improvement include using a charity as the recipient, providing more illustrative screens of the price lists in the instructions, and incentivizing correct answers to control questions. Additionally, presenting examples in the instructions based on a single switch or imposing a single-switch condition could enhance comprehension and reduce noise in the data.

Further research in this area could help provide stronger evidence on the influence of weakly informative signals on motivated beliefs. Such studies could deepen our understanding of how partial or ambiguous information, akin to fake news, can manipulate behavior, with broader implications for policy interventions aimed at combating misinformation.

Works Cited

Augenblick, N., E. Lazarus, and M. Thaler. 2021. "Overinference from Weak Signals and Underinference from Strong Signals." *arXiv Preprint* arXiv:2109.09871.

Bicchieri, C., E. Dimant, and S. Sonderegger. 2023. "It's Not a Lie If You Believe the Norm Does Not Apply: Conditional Norm-Following and Belief Distortion." *Games and Economic Behavior* 138: 321-354.

Dana, J., R. A. Weber, and J. X. Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33: 67-80.

Dunlap, R. E., and A. M. McCright. 2011. "Organized Climate Change Denial." In *The Oxford Handbook of Climate Change and Society*, 144-160. Oxford University Press.

Enke, B., and F. Zimmermann. 2019. "Correlation Neglect in Belief Formation." *The Review of Economic Studies* 86 (1): 313-332.

Exley, C. L. 2016. "Excusing Selfishness in Charitable Giving: The Role of Risk." *The Review of Economic Studies* 83 (2): 587-628.

Fischbacher, U. 2007. "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics* 10 (2): 171-178.

Garcia, T., S. Massoni, and M. C. Villeval. 2020. "Ambiguity and Excuse-Driven Behavior in Charitable Giving." *European Economic Review* 124: 103412.

Godlee, F., J. Smith, and H. Marcovitch. 2011. "Wakefield's Article Linking MMR Vaccine and Autism Was Fraudulent." *BMJ* 342.

Greiner, B. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114-125.
Kieren, Pascal, and Martin Weber. 2022. "Expectation Formation Under Uninformative Signals." *Expectation Formation Under Uninformative Signals*. SSRN.
Mesa-Vázquez, E., I. Rodriguez-Lara, and A. Urbano. 2021. "Standard vs Random Dictator Games: On the Effects of Role Uncertainty and Framing on Generosity." *Economics Letters* 206: 109981.
Michaels, D. 2008. *Doubt Is Their Product: How Industry's Assault on Science Threatens Your Health*. Oxford University Press.
Momsen, K., and M. Ohndorf. 2022. "Information Avoidance, Selective Exposure, and Fake (?) News: Theory and Experimental Evidence on Green Consumption." *Journal of Economic Psychology* 88: 102457.
O'Connor, C., and J. O. Weatherall. 2019. *The Misinformation Age: How False Beliefs Spread*. Yale University Press.
Oreskes, N., and E. M. Conway. 2011. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming*. Bloomsbury Publishing USA.
Poinas, F., J. Rosaz, and B. Roussillon. 2012. "Updating Beliefs with Imperfect Signals: Experimental Evidence." *Journal of Risk and Uncertainty* 44: 219-241.
Supran, G., and N. Oreskes. 2017. "Assessing ExxonMobil's Climate Change Communications (1977–2014)." *Environmental Research Letters* 12 (8).

Appendix

The structure of the price lists, i.e., the value of the sure amounts and the positive payoff in the lottery depend on the scenario. Sure amount for self ranges from 0 to 10 euros, with 0.5 cents increments, while the for other it ranges from 0 to an amount X euros, increments 1/X. The lottery has a null and a positive outcome with same *p*, and the positive outcome for self is 10 euros, while for other it is X euros. The amount X is determined in the first phase of the experiment preceding the choices in the price lists in the four different scenarios. In these initial phase all subjects engage in a so called calibration task to determine their individual value of X, that is the donation-equivalent to payoff for self: the subjects fill in a price list with 16 choices, where each choice consists in a assigning a safe amount to oneself or to another subject; the amount assigned to self is always 10 euros, while the amount assigned to the other subject is increasing in the price list, ranging from 2 euros to 30 euros by 2 euros increments. This calibration task ensures that participants are indifferent between the non-zero payoffs in the lotteries for self and in the lotteries for the other. Otherwise, different responses to risk in lotteries for self and lotteries for others may have resulted from participants valuing money for themselves and the other differently. See Figures 1A, 2A, 3A, 4A and 5A for an example from the software of the calibration task and the consequent 4 price lists for the four scenarios. The subject should start choosing 10 euros for self in the first choice, where the sure amount for the other is 0, and then switch at some point. The value of X is the amount assigned to the other in the choice where the switch occurs. It is possible that the choice is censored, i.e., the subjects always choose 10 for himself; in this case the value of X is 30.

Figure 1A – calibration task

This figure illustrates the calibration task. This is a screenshot from the software showing the case in which X=18.

Decisione	Opzione A	Opzione B	La tua scelta
1	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 0€.	Opzione A ● ○ Opzione B
2	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 2€.	Opzione A ● ○ Opzione B
3	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 4€.	Opzione A ● ○ Opzione B
4	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 6€.	Opzione A ● ○ Opzione B
5	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 8€.	Opzione A ● ○ Opzione B
6	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 10€.	Opzione A ● ○ Opzione B
7	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 12€.	Opzione A ● ○ Opzione B
8	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 14€.	Opzione A ● ○ Opzione B
9	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 16€.	Opzione A ● ○ Opzione B
10	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 18€.	Opzione A ○ ● Opzione B
11	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 20€.	Opzione A ○ ● Opzione B
12	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 22€.	Opzione A ○ ● Opzione B
13	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 24€.	Opzione A ○ ● Opzione B
14	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 26€.	Opzione A ○ ● Opzione B
15	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 28€.	Opzione A ○ ● Opzione B
16	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 30€.	Opzione A ○ ● Opzione B

Figure 2A – scenario 1

This figure illustrates price list of Scenario 1. This is a screenshot from the software and this price list holds for each value of X.

Decisione	Opzione A	Opzione B	La tua scelta
1	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 0€.	Opzione A ● ○ Opzione B
2	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 2€.	Opzione A ● ○ Opzione B
3	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 4€.	Opzione A ● ○ Opzione B
4	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 6€.	Opzione A ● ○ Opzione B
5	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 8€.	Opzione A ● ○ Opzione B
6	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 10€.	Opzione A ● ○ Opzione B
7	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 12€.	Opzione A ● ○ Opzione B
8	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 14€.	Opzione A ● ○ Opzione B
9	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 16€.	Opzione A ● ○ Opzione B
10	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 18€.	Opzione A ● ○ Opzione B
11	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 20€.	Opzione A ● ○ Opzione B
12	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 22€.	Opzione A ● ○ Opzione B
13	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 24€.	Opzione A ○ ● Opzione B
14	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 26€.	Opzione A ○ ● Opzione B
15	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 28€.	Opzione A ○ ● Opzione B
16	Tu: 10€. L'altro: 0€.	Tu: 0€. L'altro: 30€.	Opzione A ○ ● Opzione B

Figure 3A – scenario 2

This figure illustrates price list of Scenario 2. This is a screenshot from the software and this price list holds for a value of X equal to 18, as exemplified in Figure 1A. For different values of X, the positive payoff in Option A and the sure amounts would have been different as explained in the design section.

Decisione	Opzione A	Opzione B	La tua scelta
1	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 0.0€.	Opzione A ● ○ Opzione B
2	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 0.9€.	Opzione A ● ○ Opzione B
3	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 1.8€.	Opzione A ● ○ Opzione B
4	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 2.7€.	Opzione A ● ○ Opzione B
5	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 3.6€.	Opzione A ● ○ Opzione B
6	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 4.5€.	Opzione A ● ○ Opzione B
7	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 5.4€.	Opzione A ● ○ Opzione B
8	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 6.3€.	Opzione A ● ○ Opzione B
9	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 7.2€.	Opzione A ● ○ Opzione B
10	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 8.1€.	Opzione A ● ○ Opzione B
11	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 9.0€.	Opzione A ● ○ Opzione B
12	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 9.9€.	Opzione A ● ○ Opzione B
13	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 10.8€.	Opzione A ● ○ Opzione B
14	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 11.7€.	Opzione A ● ○ Opzione B
15	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 12.6€.	Opzione A ● ○ Opzione B
16	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 13.5€.	Opzione A ● ○ Opzione B
17	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 14.4€.	Opzione A ● ○ Opzione B
18	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 15.3€.	Opzione A ● ○ Opzione B
19	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 16.2€.	Opzione A ● ○ Opzione B
20	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 17.1€.	Opzione A ○ ● Opzione B
21	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 18.0€.	Opzione A ○ ● Opzione B

Figure 4A – scenario 3

This figure illustrates price list of Scenario 3. This is a screenshot from the software and this price list holds for a value of X equal to 18, as exemplified in Figure 1A. For different values of X, the sure amounts would have been different as explained in the design section.

Decisione	Opzione A	Opzione B	La tua scelta
1	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 0.0€.	Opzione A ● ○ Opzione B
2	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 0.9€.	Opzione A ● ○ Opzione B
3	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 1.8€.	Opzione A ● ○ Opzione B
4	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 2.7€.	Opzione A ● ○ Opzione B
5	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 3.6€.	Opzione A ● ○ Opzione B
6	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 4.5€.	Opzione A ● ○ Opzione B
7	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 5.4€.	Opzione A ● ○ Opzione B
8	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 6.3€.	Opzione A ● ○ Opzione B
9	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 7.2€.	Opzione A ● ○ Opzione B
10	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 8.1€.	Opzione A ● ○ Opzione B
11	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 9.0€.	Opzione A ● ○ Opzione B
12	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 9.9€.	Opzione A ● ○ Opzione B
13	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 10.8€.	Opzione A ● ○ Opzione B
14	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 11.7€.	Opzione A ● ○ Opzione B
15	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 12.6€.	Opzione A ● ○ Opzione B
16	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 13.5€.	Opzione A ● ○ Opzione B
17	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 14.4€.	Opzione A ● ○ Opzione B
18	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 15.3€.	Opzione A ● ○ Opzione B
19	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 16.2€.	Opzione A ○ ● Opzione B
20	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 17.1€.	Opzione A ○ ● Opzione B
21	Tu: ricevi 10€ con probabilità del 50%, altrimenti 0. L'altro: 0€.	Tu: 0€. L'altro: 18.0€.	Opzione A ○ ● Opzione B

Figure 5A – scenario 4

This figure illustrates price list of Scenario 4. This is a screenshot from the software and this price list holds for a value of X equal to 18, as exemplified in Figure 1A. For different values of X, the positive payoff in Option A would have been different as explained in the design section.

Decisione	Opzione A	Opzione B	La tua scelta
1	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0€. L'altro: 0€.	Opzione A ● ○ Opzione B
2	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 0.5€. L'altro: 0€.	Opzione A ● ○ Opzione B
3	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 1€. L'altro: 0€.	Opzione A ● ○ Opzione B
4	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 1.5€. L'altro: 0€.	Opzione A ● ○ Opzione B
5	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 2€. L'altro: 0€.	Opzione A ● ○ Opzione B
6	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 2.5€. L'altro: 0€.	Opzione A ● ○ Opzione B
7	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 3€. L'altro: 0€.	Opzione A ● ○ Opzione B
8	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 3.5€. L'altro: 0€.	Opzione A ● ○ Opzione B
9	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 4€. L'altro: 0€.	Opzione A ● ○ Opzione B
10	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 4.5€. L'altro: 0€.	Opzione A ● ○ Opzione B
11	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 5€. L'altro: 0€.	Opzione A ● ○ Opzione B
12	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 5.5€. L'altro: 0€.	Opzione A ● ○ Opzione B
13	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 6€. L'altro: 0€.	Opzione A ● ○ Opzione B
14	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 6.5€. L'altro: 0€.	Opzione A ○ ● Opzione B
15	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 7€. L'altro: 0€.	Opzione A ○ ● Opzione B
16	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 7.5€. L'altro: 0€.	Opzione A ○ ● Opzione B
17	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 8€. L'altro: 0€.	Opzione A ○ ● Opzione B
18	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 8.5€. L'altro: 0€.	Opzione A ○ ● Opzione B
19	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 9€. L'altro: 0€.	Opzione A ○ ● Opzione B
20	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 9.5€. L'altro: 0€.	Opzione A ○ ● Opzione B
21	Tu: 0€. L'altro: riceve 18€ con probabilità del 50%, altrimenti 0€.	Tu: 10€. L'altro: 0€.	Opzione A ○ ● Opzione B

DOI 10.17473/2240-7987-2024-2-13

Believe It or Not – An Empirical Study on Fake News Sharing*

Margherita Benzi**, Irene Maria Buso***, Paolo Chirico****, Jacopo Marchetti*****, Marco Novarese******, Giacomo Sillari*******

Abstract

In this paper, after an introduction (section 1) we present the two main alternative approaches to the acceptance and spread of fake news (section 2) and we focus on the problem of distinguishing between believing and sharing fake news (section 3). This problem becomes one of the main topics of our empirical study described in sections 4 and 5. In particular, we explore the ability to recognize the truth of the news published on social networks and the relationship between accuracy and the willingness to share news. An original aspect of our research is the use of a particular version of the trust game, a methodology developed in the environment of behavioral economics. With this methodology, we aim to better verify the actual dispositions of the subject to share and accept news on social networks. In the discussion of the experiment (section 6), results suggest that accuracy (ability to distinguish true from false news) is not the only factor in sharing but is flanked by factors linked to the affirmation of one's identity. Besides, data on sharing suggest that people tend to mitigate polarization by favoring inclusiveness and less polarized news.

* We thank CESARE Lab for the use of the laboratory facilities and the audience of the workshop The Democratic Containment of Fake News and Bad Beliefs held in LUISS in October 2023 for valuable comments and insights. Research realized with the contribution of the Italian Ministry of University and Research as part of the PRIN project 2017, Deceit and Self-Deception. How We Should Address Fake News and Other Cognitive Failures of the Democratic Public, 2017S4PPM4_004.

** ORCID: 0000-0003-4934-6494.

*** ORCID: 0009-0009-2828-1602.

**** ORCID: 0000-0003-4229-0440.

***** ORCID: 0000-0002-2415-0113.

****** ORCID: 0000-0001-5622-5549.

******* ORCID: 0000-0002-3243-1206.

Summary: Introduction. – II. The cognitive science of fake news. – III. Believing vs. Sharing. – IV. A study on sharing. – V. Results. – – Discussion and Conclusion. – Works Cited.

Introduction

The study of fake news fully emerged as a specific research area around 2016, following the particular use of digital platforms in the Brexit and US presidential election campaigns, gathering contributions from scholars from different disciplinary fields, such as psychology, various areas of philosophy, political and social sciences and computer science. Within this area, the so-called 'cognitive science of fake news', emerged as the study of the reasoning and behavior of news acceptance and dissemination by individuals exposed to misinformation in various ways (Levy and Ross 2021). These kinds of research typically consider fake news in a broad sense, including news fabricated with the intent to deceive, inaccurate information that circulates without any intention to mislead, or news indifferent to the truth, such as bullshit. Although we are aware of the methodological differences between experimental approaches in behavioral economy, social sciences and psychology, in this paper, we will use the term 'cognitive science of fake news', to refer to research grounded on experimental and quantitative studies.

The considerable amount of empirical and experimental work on misinformation imposes itself overwhelmingly on philosophical work that is not only theoretical or conceptual: political philosophers, philosophers of language and epistemologists – especially those working in the field of social epistemology – are compelled to confront empirical evidence. One way of doing this is by working with scholars from other disciplines, such as statisticians and social scientists. The purpose of this paper is twofold: on the one hand, it intends to present a contribution to research on why people believe in fake news and spread it; on the other hand, to indicate some of the methodological and epistemological aspects that emerged in the course of work conducted by an interdisciplinary team.

II. The cognitive science of fake news

We will begin by briefly recalling the two main explanatory hypotheses concerning the acceptance and spread of fake news. The first, sometimes

called the *bounded rationality* approach, or "cognitive account", builds on a pre-existing strand of research that investigated conspiracy theories, self-deception, and the acceptance of pseudoscientific theories, especially in the medical field. A large part of this research, inspired by the *Heuristics and Biases* approach (Tversky and Kahneman 1974), was focused on investigating the cognitive distortions that make us inclined to adopt epistemically suspect beliefs and behave accordingly. Within this theoretical framework, the acceptance and transmission of false or inadequate information would depend on 'cognitive vices' such as the incapacity to distinguish truthful from untruthful news, or on reasoning guided by the fast, intuitive, automatic, and emotional cognitive mechanisms that make up system 1 (Stanovich and West 2000, Kahneman 2003), as distinct from slow, deliberative, analytical, rational, and logical thinking (system 2). According to this approach, therefore, falling prey to misinformation depends on ignorance, and consequent insufficient ability to distinguish truthful from untruthful information, or on inferential capacities limited by contingencies, such as distraction, and poor analytical skills.

The second explanatory hypothesis, which has been established since the 2000s in the field of political polarization and motivated reasoning research, is *expressive rationality*. It grounds on the thesis that, motivated to safeguard and uphold beliefs that are crucial to their identities, individuals tend to accept information that aligns with their ideology while rejecting information that does not (see Van Bavel and Pereira 2018). Particularly influential on this issue was an article by Kahan (2013) strongly critical of the thesis that motivated reasoning – understood as "the tendency of people to conform assessments of information to some goal or end extrinsic to accuracy"- is linked to errors in reasoning. The work presented experimental evidence suggesting that motivated reasoning (particularly that of conservatives, who constituted the study population) does not arise from reasoning shortcuts like heuristics. On the contrary, it was more frequent in subjects who scored higher on tests of analytical reasoning ability. Based on these findings, Kahan presented his alternative hypothesis/theory to the strictly cognitivist one: recourse to 'motivated thinking' activates information processing procedures that do not aim at the truth but are nevertheless advantageous for the individuals since maintaining such beliefs expresses their loyalty to a group of people with which they have positive relations. Saying something false about subjects outside one's expertise does not generally entail great penalties; on the contrary, dissenting from one's group incurs high costs, affective, moral, and material. Therefore, if motivated reasoning promotes the interests of those who resort to it, it cannot, for Kahan, be considered irrational *tout-court*: rather, it follows what, according

to the author, could be called "expressive rationality". In support of his hypothesis, Kahan (2013, 2015) presented empirical evidence suggesting that the most polarized people scored well on the Cognitive Reflection Test ("CRT"), and even scored higher than others on the Ordinary Science Intelligence ("OSI") test, which measures both basic scientific factual knowledge and reasoning skills. Kahan's hypothesis would prove decisive in posing what still today is one of the central questions in the theoretical debate on fake news: do we believe fake news, and spread it, because we are subject to some 'defect' in our inferential procedures, or because we are protecting, or affirming, our identity? Studies that illustrate the two positions include Bronstein et al. (2019), and Pennycook and Rand (2021) in defense of the bounded rationality hypothesis, and Osmundsen et al. (2021) in favor of expressive rationality.

On a very general level, it can be said that much of the research on fake news is based on two why-questions:

(i) Why do people believe *fake* news to be *true*?
(ii) Why do they contribute to its spread?

Given its relevance in different fields, the problem of the acceptance/dissemination of fake news has been tackled using a variety of methods. In a review of 202 articles on fake news on the Internet published between 2018 and 2021 in 117 journals (mainly from the social sciences, psychology and communication studies), Wu et al. (2022) identify seven research methodologies: conceptual approach, case studies, focus group interviews, surveys, experiments, secondary data analysis (i.e., use of data that was collected by someone else for some other purpose) and modelling and simulations.

The methodologies indicated by Wu et al. are not mutually exclusive. Conceptual analysis, along with appropriately collected and analyzed data, can inform a theory (or model), which can be subsequently evaluated by experimentation. Alternatively, a survey can be used as a preliminary stage in the building of a causal model, to be tested with subsequent experiments or surveys. The use of experiments and surveys to test causal hypotheses has a well-established tradition in cognitive psychology and has exhibited an increasing trend in sociology, political science, and economics over the last thirty years, thanks to the 'causal revolution' (Mize and Manago 2022, Pearl 2009) that took hold in the social sciences in the first decades of the century.

Concerning the two prevailing research agendas in the cognitive science of fake news (bounded rationality vs. expressive rationality), the comparison is

mainly made through experiments aimed at confirming the basic explanatory hypothesis (i) and (ii) by refuting rival hypotheses. Schematically, an ideal causal experiment consists of the controlled variation of a variable while holding fixed, through random assignment, all other factors that could be causally relevant for the presumed effect. As Mize and Manago point out, these two elements, the manipulation of an independent variable and the random assignment of subjects, are common to causal experiments in all social sciences. Taken together, they increase the plausibility that changes in the variable representing the presumed effect (the dependent variable) are generated by changes in the independent variable rather than by other possible causes (2022, p. 102799). From this perspective, however, considering all the factors that could be causally relevant to the effect makes the design of experiments extremely complex. Tappin, Pennycook and Rand (2020) criticize some current uses of experiments in fake news research, stressing that poorly designed experiments can undermine the correctness of causal inference. By way of example, the two authors criticize two types of experiments (called *outcome switching* and *party cues*, respectively) that due to their preponderance in the literature on politically motivated reasoning (the authors count 51 experiments of one or the other type in 40 years of research in psychology) can be called paradigmatic. In fake news research, the causal hypothesis submitted to the experimental control was, in summary, "Political motivation directly determines the judgement on the reliability of the news." However, even when the outcome of the experiments confirmed this hypothesis, the evidence provided by the experiments was, according to Pennycook and Rand, of little relevance, as they merely tested the direct causal relationship between political motivation and reasoning, but did not satisfy the *excludability assumption*, according to which the treatment – in this case political motivation – is the *only* causally relevant factor. In both types of paradigmatic experiments, it is possible to postulate other causal pathways, possibly with additional variables that influence the outcome and are external to the *direct* pathway 'Political Motivation → Reasoning'. Although Pennycook and Rand's polemical target was the expressive rationality hypothesis, their methodological remarks have general validity. Even in its most sophisticated formal versions, causal inference requires considerable theoretical commitment in the choice of factors that might, and can, influence the phenomenon one wishes to explain, and in the possible causal pathways that unite them. In this framework, surveys, such as the one presented in this study, are useful because they often bring out statistical relationships that are worth investigating from a causal perspective.

III. Believing vs. Sharing

Most of the contributions mentioned in the preceding session focus on the first question: Why do we believe in fake news? The second question – Why do we help spread them? – has recently gained increasing independent interest, particularly concerning information-sharing behavior via social media (see Melchior and Oliveira 2024). Many studies (e.g. Acerbi 2019; Bor et al. 2020; Altay et al. 2023) show a considerable discrepancy between news that is *believed* to be true and news that is *shared,* stressing that "sharing is not believing". The explanation of this discrepancy has implications for the theoretical debate referred to in the previous paragraphs, as a major split between believing fake news to be true and sharing it on social media seems to support the expressive rationality approach and assumptions about the motivations for dissemination other than purely informative and accuracy-based ones. Pennycook and Rand (2021) and Pennycook et al. (2021), present data confirming the existence of a split between the preference for truthful news and the intention to share it. The authors, however, deny that this split is (entirely) due to the fact that the intention to share is mainly driven by polarization and/or assertion of political affiliation, and argue that it is largely due to distraction:

> "Finally – and most consistent with the earlier focus on a lack of reflective thinking being a source of misjudgments – the inattention-based account argues that people have a strong preference to only share accurate content, but that the social media context distracts them from this preference. Consistent with this account, asking participants to rate the accuracy of each headline before deciding whether to share it decreased sharing of false headlines by 51% relative to the baseline condition [...] – suggesting that inattention to accuracy was responsible for roughly half of the misinformation sharing in the experiment." (Pennycook and Rand 2021, 395)

In contrast, Osmundsen et al. (2021) deny that the evidence in the literature and the debate in psychology on online trolling do not support the hypothesis that the spread of fake news is mainly related to distraction or ignorance and add to the pool of possible individual motivations the so-called *disruption theory*. The latter is based on the propensity for trolling behavior, seen as "pleasure in misleading and harassing others online, and political cynicism, understood as a generic anti-system or anti-elite attitude" (p. 1002).

The experimental part of Osmundsen et al.'s work confirms, besides the relevance of polarization, some aspects that had already emerged in the exper-

imental literature, such as the greater occurrence of spreading fake news in older age groups, even it was not associated with greater ignorance. The study does not explicitly distinguish between conscious dissemination and the spread of fake news, although the consideration of trolling and political cynicism predicts dissemination attitudes without belief. A direct investigation of conscious dissemination of fake news was done by Chadwick, Vaccari and O'Loughlin, who, in their investigation of 'dysfunctional sharing' of information, identified three possible explanatory variables that could motivate conscious dissemination of fake news: i) *persuading/informing*, ii) *debating*, and iii) *entertaining/trolling*, a term used for "the clustered motivations to entertain, please, and upset others" (2018, 4264):

> "Trolling is a contested and context-dependent term that captures multiple practices, some of which are positive [...]. Nevertheless, let's think about a continuum of positive to negative online political behaviour. We suggest trolling is comparatively negative, and it is all the more important to be clear about this in a so-called 'post-truth' context." (*ibid.*, 4264).

In a subsequent qualitative-quantitative study, Ardevol-Abreu, Delponti and Rodrıguez-Wanguemert (2020) wanted to expand on the research of Chadwick et al. (2018), focusing, among other things, on the motivations for consciously spreading fake news. When interviewed in this regard, some people invoked freedom of expression, stating that they agreed with discourses that they knew to be wrong (and thus showing, according to the authors, that they were unable to distinguish between facts and opinions). Others said that they spread them with debunking intentions, accompanying them with a critical commentary, and finally, others, especially in the area of health information, claimed that they shared some news that, although not sufficiently verified, nevertheless expressed possible assertions, which could change their truth value from false to true over time. Regarding the characteristics of the conscious spreaders of fake news, the authors examined the information provided by the subject survey depending on gender, age, education, strength of ideological involvement, trust in the government, and concern about the COVID-19 epidemic, without finding significant differences between the subgroup of conscious spreaders of fake news and that of the other respondents.

The case of people who spread news believing it to be true also has interesting aspects. Which news, among those believed to be true, is most likely to be shared, and why? t'Sersteven, Piccillo and Grigoriev (2022) define 'activist behavior' as a behavior in which the higher perceived veracity of a claim leads

to increased reaction likelihood and argue that the dissemination of fake news is mainly motivated by this type of attitude. Shephard et al. (2023) present an experimental study with Scottish participants on the relationship between sharing behavior and content. In a range of topics that fell into seven categories (Crime, Economy, Education, Europe, Scotland, Health, and Immigration), health-related news appeared to be the most shared, whereas immigration was the least shared. Again, the authors leaned towards an explanation of sharing based on the 'activist' propensity to share useful information:

> "That dichotomy is between users who share fake news because it fits with their worldview (i.e., sharing for doctrinal reasons), and 'best of intentions' users, those individuals who feel the need to share important content that could be useful to others in their networks. This is manifested in the high sharing rates for real and fake health news, in which both types of users combine to increase sharing rates (i.e., doctrinal; I knew I was right about that issue and 'best of intentions': my network need to know about this). In contrast, the low sharing rates for immigration might be explained by that content only appealing to users with a specific doctrinal interest in such news." (Shephard et al. 2023, 8)

The relationships between exposure, truthfulness assessment, and sharing appear particularly relevant, not only because they can help settle the theoretical controversy on the respective role of cognitive weakness or expressive rationality, but also for a correct assessment of the 'democratic dysfunctionality' of the transmission of information on social platforms, and the possible remedies to be adopted. While being aware of the damage produced by misinformation in the political sphere, one should not, as Altay et al. (2023) remind us, take it for granted that the attribution of 'likes' corresponds to an endorsement of the news: "Sharing and liking are not believing. People interact with misinformation for a variety of reasons: to socialize, to express skepticism, outrage or anger, to signal group membership [...] or simply to have a good laugh" (5). Even less should it be assumed without further investigation that the belief that certain news is true implies changes in behavior (Altay et al. 2023). One side of this debate posits that while fake news may contribute to disinformation, it does not alter attitudes or voting behavior. This is because individuals are inclined to accept false information that aligns with their preexisting partisan preferences (see for instance Van Bavel and Pereira 2018). In this regard (Galeotti 2021) focuses on voting behavior. She argues that if the main source of fake news is motivated reasoning, and more specifically "that form of motivated reasoning induced by ideological beliefs and partisan affilia-

tions" (64), and if partisan affiliations largely drive the dissemination of information on social media, political misinformation will reinforce pre-existing beliefs in the case of favorable content or be rejected in the case of unfavorable content. Therefore, the diffusion of fake news is unlikely to produce disruptive changes in electoral behavior. This kind of consideration prompted the empirical research we present in the next section.

IV. A study on sharing

The project 'Gathering Opinions on News from the Internet and Social Networks' implemented by scholars from different areas (political philosophy, philosophy of science, economics, statistics), is a preliminary study, specific to the Italian reality, on the behavior of accepting and sharing news on social network platforms. In particular, it intends to explore the ability to recognize the truth of the news published on social networks, and the relationship between accuracy and the willingness to share news, making a comparison with specifically social aspects, such as popularity, ties with the community, and trust. A further research question is to check whether belief and dissemination behaviors are uniform or varied according to the type of news. The study is not intended to test a causal hypothesis directly, but rather to bring out any statistical correlations to be tested for causality in possible future experiments. The project has two main aspects of originality. The first is that the survey, conducted on a digital platform, involves the general behavior of Italians: literature is scarce on this topic since most of the work on the online transmission of news by Italians concerns specific arguments, such as health information and electoral behavior. The second is that our research, as we shall see, investigates the sharing behavior of Italians through particular versions of the trust game.

The study consists of a survey administered on the digital platform Prolific [1] in March 2024 and involved 105 participants of Italian nationality, 59 males, and 46 females. The participants, paid for their participation, were divided into five age groups (Table 1).

[1] See www.prolific.com; see also Palan and Schitter (2018).

Table 1

	Gender		
Age	**Female**	**Male**	**Total**
20-29	54.3%	39.0%	45.7%
30-39	19.6%	40.7%	31.4%
40-49	10.9%	16.9%	14.3%
50-59	8.7%	0.0%	3.8%
60-70	6.5%	3.4%	4.8%
Total	**100.0%**	**100.0%**	**100.0%**

Participants were asked to specify ethnicity (whites 96%, others 3%, mixed 1%), and occupation (see Table 2 below).

Table 2

Employment	**Counting ID**
Due	2.9%
Full-Time	33.3%
Not in paid work (e.g. homemaker', 'retired or disabled)	8.6%
Other	14.3%
Part-Time	23.8%
Unemployed	10.5%
Unemployed (and job seeking)	1.0%
Not responding	5.7%
Total	**100.0%**

A second set of questions, which we report below, concerned the use of social media.

– Which social networks do you use? You can select one or more options (Facebook / X (Twitter) / Instagram / TikTok/ (Other – indicate name(s))

– How much time do you spend on social networks? (1 hour or more per day / less than 1 hour per day)

– Do you use social networks for leisure or work? (Mostly for leisure / Mostly for work / Both for leisure and work)

– Do you use social networks to inform yourself about current news? (Yes / No)

The questions in the third group were about sixteen news stories, true or false, on various topics concerning current affairs, politics, economics, and society, mostly found on fact-checking sites (AGI, *Pagella Politica*). Below are the titles of the news stories:

1. Data leave no doubt: too few public servants in Italy. [True]

2. Iraqi executed by Isis; here's why he wore a Napoli jersey. [False]

3. Disabled person gets stuck in City Hall elevator for two hours: fined for screaming. [False]

4. The Padua prosecutor's office challenges 33 birth certificates of babies with two mothers. "They are illegitimate." [True]

5. Migrants, Salvini's iron fist. "Germany pays NGOs to bring them. Don't rule out the use of the navy." [True]

6. The statement, "Roma camps do not exist anywhere else in Europe, I don't see why they should exist in Italy." [False]

7. Public investment spending in the South: more than halved in ten years. [True]

8. Soros and MasterCard, apparently in cooperation with UNICHR and EU, made a deal to equip immigrants with credit cards to stimulate their consumption as they wander around Europe. [False]

9. Greenhouse gases, ISPRA: emissions down 19% in the 30 years. [True]

10. Toxic GMO mosquitoes: new weapons of war? [False]

11. Immigrants in Italy generate 9% of the Gross Domestic Product. [True]

12. Giorgia Meloni: The tragedy of Acca Larenzia cannot be a pretext for nostalgic demonstrations that a modern Right clearly shuns. [False]

13. Vaccine Dictatorship: the EU Call for Coercion. [False]

14. The Supreme Court has ruled that the fascist salute in its expression is not a crime. [True]

15. Mandatory doggy bag in restaurants. [True]

16. Artificial intelligence chatbots are politically aligned. [True]

For each news item, participants were asked eight dichotomous (yes / no) questions:

(i) [*Think True*] In your opinion, is this news true or false?

(ii) [*Believe True*] Do you think people using your social networks believe this news is true?

(iii) [*Share*] Would you share this news on the social networks you usually use?

(iv) [*Likes/People*] In your opinion, would this news get many "likes" from people who use your social networks?

(v) [*Likes/Contacts*] Do you think this news would get many "likes" from your contacts if you posted it on your social networks?

Questions(iv) and (v) are designed to see if there is any perceived difference between one's friends (v) and the rest of the world (iv). The last three questions are based on particular versions of the Trust game:

(vi) [*Trust*] In the first game, which we will call "Trust", we ask, for each news item, whether the participant would choose that news item to present her/himself to a partner – unknown and randomly assigned – who must decide whether to donate money, and how much, to the participant.

(vii) [*Network*] The second game, which we will call "Network", is a variation of the first one, where, however, the pairings are not random: the player is matched with a partner who decides whether to accept him into his or her group depending on the news published by the player. After eventual acceptance, the "Trust" game is played with the people in the group thus formed. For each of the 16 news items, participants are asked if they would use it to introduce themselves to the group by which they would like to be accepted.

(viii) [*Accept*] The third and final game, called "Accept", is the counterpart of the second one: participants are asked, for each news item, whether they would agree to include in their group people who post that news item in their social profile (and play a "Trust-game" with them). The explicit question is in this case, "Would you accept into your group a person who posted this news on his or her profile on the social network?" The assumption is that if the person answers in the affirmative, he or she trusts the person who posted the news.

V. Results

Table 3 below shows the sums of the responses obtained in the Survey to the eight questions out of the 16 news items.

Table 3

NEWS	Think True	Believe True	Share	Like/ People	Like/ Contacts	Trust	Network	Accept	**total**
1	61	83	12	35	15	14	17	51	288
2	53	75	7	53	29	10	10	27	264
3	18	55	9	59	32	9	11	26	219
4	76	89	32	61	35	32	31	48	404
5	80	92	15	64	27	8	11	17	314
6	50	89	7	59	22	11	8	15	261
7	82	96	44	57	40	42	36	57	454
8	6	49	0	42	8	1	1	5	112
9	56	68	32	43	33	39	39	54	364
10	8	57	4	49	15	4	3	13	153
11	92	52	42	33	36	41	41	64	401
12	55	81	4	44	12	8	6	17	227
13	14	68	4	51	13	3	3	10	166
14	62	85	18	56	20	9	8	17	275
15	75	87	44	67	45	46	42	54	460
16	31	76	16	49	20	17	15	29	253
Total	**819**	**1202**	**290**	**822**	**402**	**294**	**282**	**504**	**4615**

Data processing provided information on three aspects relevant to the original questions: the ability to evaluate the truth of the news, sharing behaviors, and possible associations between truth-value recognition, sharing, and trust. Overall, subjects showed good recognition ability: true news was recognized in 65% of cases (almost two out of three) and believed false in 35%; fake news was recognized as such in 72% and believed true in 28% of cases. Among the true news, those with the highest identification index were: 11 – *Migrants and GDP* (88%), 7 – *South Spending* (78%), and 5 – *Navy vs. NGOs* (76%). Among the true news, those most frequently found to be false were: 16 – *Doggy Bag* (70% errors), 9 – *Greenhouse Gas* (47%), and 1 – *Public Servants* (42%). Among the fake news, those with the highest recognition rates (i.e. correctly identified as fake) were: l8 – *Soros and MasterCard*, 10 – *Toxic GMO Mosquitoes,* and 3 – *Fine to Disabled*, with correct recognition rates of 94%, 92%, and 86%, respectively. Conversely, about half of the participants mistakenly judged as true news items 12 – *Meloni*, 2 – *Iraqi Executed*, and 6 – *Roma Camps*, with error rates of 52%, 50%, and 48%, respectively.

An interesting aspect emerges when comparing the answers to the questions "In your opinion, is this news true or false?" and the next one, "Do you think people using your social networks believe this news is true?" A comparison of the first two columns of Table 3) shows that in many cases people

think that the news is not true but yet others believe it to be so; thus, there seems to emerge a tendency to consider themselves more capable than the average user of their networks in judging the accuracy of the information.

The survey produced partially unexpected results regarding the relationship between accuracy and sharing behavior. The study showed a low propensity to share: only 33% of news deemed true and 3% of news deemed false are shared. Shared news is true and believed to be true in 82% of cases and false but believed to be true in 10% of cases. Accuracy is therefore a relevant factor: believing the news to be true appears to have a strong influence on its dissemination. However, believing a piece of news to be true is not a sufficient requirement for sharing (although almost necessary): other factors seem to play an important role. The table shows, for each piece of news, how many people believe it to be true, the percentages of people who would be willing to share it, how many would use it in *Trust*, how many in *Network*, and how many would be willing to *Accept* a person sharing it (a proxy for the trust assigned to that person).

Table 4

	think true		Among people who think the news true			
News	**(%)**	**num.**	**Share**	**Trust**	**Network**	**Accept**
1	58.1%	61	18.0%	16.4%	21.3%	63.9%
2	50.5%	53	13.2%	13.2%	15.1%	37.7%
3	17.1%	18	33.3%	27.8%	33.3%	44.4%
4	72.4%	76	40.8%	42.1%	40.8%	59.2%
5	76.2%	80	17.5%	10.0%	13.8%	21.3%
6	47.6%	50	12.0%	12.0%	10.0%	18.0%
7	78.1%	82	53.7%	50.0%	42.7%	63.4%
8	5.7%	6	0.0%	0.0%	0.0%	16.7%
9	53.3%	56	53.6%	62.5%	58.9%	75.0%
10	7.6%	8	50.0%	37.5%	37.5%	75.0%
11	87.6%	92	44.6%	43.5%	43.5%	65.2%
12	52.4%	55	7.3%	12.7%	9.1%	29.1%
13	13.3%	14	21.4%	14.3%	14.3%	50.0%
14	59.0%	62	25.8%	12.9%	11.3%	21.0%
15	71.4%	75	50.7%	54.7%	49.3%	60.0%
16	29.5%	31	38.7%	35.5%	41.9%	58.1%
Total	**48.8%**		**32.6%**	**31.3%**	**30.4%**	**48.6%**

Table (4) shows some discrepancies between accuracy and acceptance. In descending order, the news with an acceptance value as true higher than average are: 11- *Migrants and GDP*, 7- *South Spending*, 5- *Navy vs. NGOs*, 4- *Two*

Mothers, 15- *Doggy Bag*, 14- *Fascist Salute*, 1- *Public Servants*, 9- *Greenhouse Gas*, 2- *Iraqi executed*, 12- *Meloni*, (about 49%). The least believed news is 8- *Soros and MasterCard*. Amongst these, news 7, 9, 11, 4, and 15 also appear, albeit in a different order, among the ten news items one is most willing to share; however, this ranking also sees news 10- *Toxic GMO Mosquitoes*, 16- *Chatbot*, 3- *Fine to Disabled*, followed by 14- *Fascist Salute* and 13- *Vaccination Dictatorship*. News item 8 appears again to be the least agreeable (0%).

A similar reading of the data referring to the news considered to be false shows an even more pronounced mismatch between accuracy and sharing. Table 5 below shows, for each piece of news, how many people think it is false and the percentages of people who would be willing to share it, how many would use it in the *Trust*, how many in the *Network*, and how many would be willing to *Accept* a person sharing it.

Table 5

	Think False		Among those who think the News is false			
News	**(%)**	**num.**	**Share**	**Trust**	**Network**	**Accept**
1	41.9%	44	2.3%	9.1%	9.1%	27.3%
2	49.5%	52	0.0%	5.8%	3.8%	13.5%
3	82.9%	87	3.4%	4.6%	5.7%	20.7%
4	27.6%	29	3.4%	0.0%	0.0%	10.3%
5	23.8%	25	4.0%	0.0%	0.0%	0.0%
6	52.4%	55	1.8%	9.1%	5.5%	10.9%
7	21.9%	23	0.0%	4.3%	4.3%	21.7%
8	94.3%	99	0.0%	1.0%	1.0%	4.0%
9	46.7%	49	4.1%	8.2%	12.2%	24.5%
10	92.4%	97	0.0%	1.0%	0.0%	7.2%
11	12.4%	13	7.7%	7.7%	7.7%	30.8%
12	47.6%	*50*	0.0%	2.0%	2.0%	2.0%
13	86.7%	91	1.1%	1.1%	1.1%	3.3%
14	41.0%	43	4.7%	2.3%	2.3%	9.3%
15	28.6%	30	20.0%	16.7%	16.7%	30.0%
16	70.5%	74	5.4%	8.1%	2.7%	14.9%
Total	**51.3%**		**2.7%**	**4.4%**	**3.8%**	**12.3%**

The most shared news items (i.e., above average) are, in descending order: 15, 11, 16, 14, 9, 5, 3 and 4. Of these, 4, 5, 9, 11, 14 and 15 are also present in the most popular news list among those believed to be true; 8 has value 0. It therefore appears that some news items are highly disseminated regardless of the accuracy attributed to them: in particular 11, 9, 4 and 15.

Let us now examine the relationships between truth, sharing and trust. Trust is relevant to the relationship between truth and sharing because discovering that the demand or concession of trust is based on accuracy would favor the bounded rationality approach, whereas discovering that trust depends on identity aspects would favor the expressive rationality approach. It should be noted, however, that in the transition from *Trust* to *Network* we see a change in the communicational environment: in the former case one knows nothing about the partner, and it is important to accredit oneself as a trustworthy person as opposed to a generic partner: this could lead to a certain caution in expressing controversial opinions. On the contrary, in *Network* one chooses the partners and therefore group identity takes on importance: this could favor identity motivations. Consequently, the magnitude of the difference between the behaviors exhibited in *Trust* and *Network* respectively indicates the respective weight of accuracy and expressive usefulness. Among those who believed the news to be true, the preferred piece of news to be credited in *Trust* is 9 (with a slightly higher score than *Network*), and among those who believed it to be false is 15 (with equal scores between *Trust* and *Network*).

The other news items mostly shared by those who believe them to be fake are often different from the kind of news posted by those who believe it to be true. Specifically, in games *Trust* and *Network,* News 4 drops dramatically compared to the sharing among those who believe it to be false and increases among those who believe it to be true. News 16 increases in *Trust* and decreases in *Network* among those who believe it to be false and follows the opposite trend among those who believe it to be true.

A further remark is that people seem to favor sharing news they believe to be true, with a particular preference for news of a social, economic and environmental nature, often supported by quantitative data. For instance, see the difference between news 5 (*Migrants and Germany*) and 9 (*Greenhouse Gas*): 5 has a higher attributed truthfulness (76%) than 9 (53%), but a much lower willingness to be shared in *Share/Trust/Network* than 9. Here, we eventually address the significance of polarization and identity-defined beliefs concerning partner choice.

In general, if we define ‘polarizing’ as news that expresses a very clearly identifiable or extreme political position and is highly divisive, we can classify the news as follows:

POLARISING NEWS

News	News Topic	Polarizing
1	Public Servants	No
2	Iraqi Executed	Yes
3	Fine to Disabled	No
4	Two Mothers	No
5	Migrants and Germany	Yes
6	Roma Camps	Yes
7	South Spending	No
8	Soros and MasterCard	Yes
9	Greenhouse Gas	No
10	Toxic GMO Mosquitos	No
11	Migrants and GDP	No
12	Meloni	Yes
13	Vaccine Dictatorship	Yes
14	Fascist Salute	Yes
15	Doggy Bag	No
16	Chatbots	No

The disaggregated data (*Share*, *Trust*, *Game* and *Accept*) according to this classification are shown in the following tables.

Table 6

Polarizing	Share		
	No	Yes	Total
No	75.1%	24.9%	100.0%
Yes	92.5%	7.5%	100.0%
Total	**82.7%**	**17.3%**	100.0%

Table 7

Polarizing	Trust		
	No	Yes	Total
No	74.2%	25.8%	100.0%
Yes	93.2%	6.8%	100.0%
Total	**82.5%**	**17.5%**	100.0%

Table 8

Polarizing	Network		
	No	Yes	Total
No	75.1%	24.9%	100.0%
Yes	93.6%	6.4%	100.0%
Total	**83.2%**	**16.8%**	100.0%

Table 9

Polarizing	Accept		
	No	Yes	Total
No	58.1%	41.9%	100.0%
Yes	85.3%	14.7%	100.0%
Total	**70.0%**	30.0%	100.0%

According to the above classification of news, Tables 6 – 9 show that the percentage of sharing/use in games is reduced to about one-third in the case of polarizing news.

Further confirmation of a cautious attitude towards polarizing news comes from Tables 4 and 5. Particularly striking is the data on acceptance, which presents an average value of 48.6% in (Table 4 above), showing that almost half of the respondents who think the news is true do not accept people who present themselves by posting that news. In particular, people who show up by posting news item 5 and news item 14 – news items that are very politically characterized – receive a very low acceptance rate even from those who believe these news items to be true.

If we turn to the game concerning the acceptance in one's network of peo-

ple who present themselves by posting fake news, we see that Table (5) above shows a 'tolerant' attitude: the last column *Accept* has generally higher values than all the others; on average, 12.3% of the participants would accept to be part of a network of someone who shares fake news. In particular, tolerance emerges for those who present false but 'inclusive' news such as 11, 15, 1, and 3. In the case of news item 8, a very low percentage believe it to be true, no one shares it or uses it in *Trust* and *Network*, but 16.7% would still accept a person sharing it (this is also the lowest percentage of acceptance).

These considerations suggest that accuracy (believing a piece of news to be true) undoubtedly plays a role but is neither a sufficient nor a necessary condition for sharing.

Discussion and Conclusion

The aim of this work was to explore the ability to recognize the truth of the news published on social networks and the relationship between accuracy and disposition to share; furthermore, we wanted to investigate whether the sharing behavior was uniform or varied according to the type of news.

The analysis of data shows a satisfactory ability to distinguish true from false news and suggests that accuracy while having a significant weight, is not the only factor in sharing, but is flanked by factors linked to the affirmation of one's identity. However, this affirmation does not coincide with a group identity: on the contrary, the data on sharing suggests that people prefer to mitigate polarization by favoring inclusiveness and 'morally correct' news (on the environment, the economy, etc.). The attitude of inclusiveness and tolerance appears particularly pronounced in the *Accept* game with the relatively high percentage of acceptance in one's network of people who share fake news. A possible explanation for this behavior could be the reluctance of people to present themselves as hyper-partisan when interacting on a neutral platform. This result would be in line with similar findings in both the fake news literature (see Shephard et al. 2023), [2] as well as that of partner choice (see Martin, Young and McAuliffe 2019). In the latter perspective, presenting oneself to a

[2] "This suggests an interesting degree of dissociation in which people are engaging in some form of self-censorship in terms of what topics, or items, they would feel comfortable sharing. For Immigration news in particular, it suggests that misinformation attacks in relation to this topic, may only be successful if it is advertized to, and promoted by, those who already share strong doctrinal feelings about it." (Shephard et al 2023,4).

general audience as generous and cooperative could perhaps count for more than accuracy and bias.

From the analysis of the data, some limits concerning methodological aspects have emerged which could be considered in further work. Besides some methodological aspects such as those concerning the size of the sample, news format, news order, pictures and graphics, some other aspects should be further investigated. The classification used was not completely satisfactory. For instance, there are doubts as to whether news item 4 (*two mothers*) was polarizing or not: on the one hand it can be argued that it is polarizing; on the other hand, in this case, polarization would be across the right/left opposition. A more detailed classification of sharing motivations would also be needed (see e.g. Melchior and Oliveira 2024, who consider 26 categories of news). Finally, the binary questions did not allow for the expression of intermediate belief values between true and false, an important aspect to distinguish between the conscious spreading of falsehoods for recreational or disruptive purposes (entertaining or anti-social behavior) and the spreading of news about which one is not certain, but which could be useful (prosocial behavior). These limitations should be considered in a future experimental study.

Works Cited

Acerbi, Alberto 2019. *Cultural Evolution in the Digital Age*. Oxford: Oxford University Press.

Altay, Sacha, Manon Berriche, and Alberto Acerbi. 2023. "Misinformation on misinformation: Conceptual and methodological challenges." *Social Media + Society*, 9(1). DOI: 10.1177/20563051221150412.

Ardèvol-Abreu, Alberto, Patricia Delponti, and Carmen Rodríguez-Wangüemert. 2020. "Intentional or Inadvertent Fake News Sharing? Fact-Checking Warnings and Users' Interaction with Social Media Content." *Profesional de la Informacion* 29 (5): e290507.

Bor, Alexander, Mathias Osmundsen, Stig Hebbelstrup, Rye Rasmussen, Anja Bechmann, and Michael Bang Petersen. (2020). "'Fact-checking' videos improve truth discernment ability but do not reduce fake news sharing on Twitter." *PsyArXiv Preprint.*

Bronstein, Michael V., Gordon Pennycook, Adam Bear, David G. Rand, and Tyrone D. Cannon. 2019. "Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking," *Journal of Applied Research in Memory and Cognition* 8 (1): 108-117.

Chadwick, Andrew, Cristian Vaccari, and Ben O'Loughlin. 2018. "Do tabloids poi-

son the well of social media? Explaining democratically dysfunctional news sharing." *New media & society* 20 (11): 4255-4274.
Galeotti, Anna E. 2021. "Political disinformation and voting behavior: Fake news and motivated reasoning." *Notizie di Politeia* 37 (142): 64-85.
Grunewald, Andreas, Victor Klockmann, Alicia von Schenk, and Ferdinand von Siemens. 2024. "Are biases contagious? The influence of communication on motivated beliefs." *Würzburg Economic Papers*, 109, University of Würzburg, Department of Economics, Würzburg.
Kahan, Dan M. 2013. "Ideology, motivated reasoning, and cognitive reflection." *Judgment and Decision making* 8 (4): 407-424.
Kahan, Dan M. 2015. "Climate-Science Communication and the Measurement Problem." *Political Psychology* 36: 1-43.
Kahneman, Daniel. 2003. "Maps of bounded rationality: Psychology for behavioral economics." *American Economic Review* 93 (5): 1449-1475.
Levy, Neill and Robert M. Ross. (2021). "The cognitive science of fake news. In *The Routledge Handbook of Political Epistemology*, edited by M. Hannon and J. de Ridde (pp. 181–191). London: Routledge.
Martin, Justin W., Liane Young, and Katherine McAuliffe. 2019. "The psychology of partner choice." *PsyArXiv Preprints.*
Melchior, Cristiane, and Mírian Oliveira. 2024. "A systematic literature review of the motivations to share fake news on social media platforms and how to fight them." *New Media & Society* 26 (2): 1127-1150.
Mize, Trenton D., and Bianca Manago. 2022. "The past, present, and future of experimental methods in the social sciences." *Social Science Research* 108: 102799.
Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. "Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter." *American Political Science Review* 115 (3): 999-1015.
Pearl, Judea. 2009. *Causality*. Cambridge: Cambridge University Press.
Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. (2021). "Shifting attention to accuracy can reduce misinformation online." *Nature 592*(7855): 590-595.
Pennycook, Gordon and David G. Rand. 2021. "The psychology of fake news." *Trends in cognitive sciences* 25 (5): 388-402.
Shephard, Mark P., David J. Robertson, Narisong Huhe, and Anthony Anderson. 2023. "Everyday Non-partisan fake news: Sharing behavior, platform specificity, and detection." *Frontiers in Psychology* 14: 1118407.
Stanovich, Keith E., and Richard F. West. 2000. "Individual differences in reasoning: Implications for the rationality debate?" *Behavioral and Brain Sciences* 23 (5): 645-665.
Tappin, Ben M., Gordon Pennycook, and David G. Rand. 2020. "Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic

study designs often undermine causal inference." *Current Opinion in Behavioral Sciences* 34: 81-87.
T'Serstevens, François, Giulia Piccillo, and Alexander Grigoriev. 2022. "Fake news zealots: Effect of perception of news." *Frontiers in psychology* 13: 859534.
Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *Science* 185 (4157): 1124-1131.
Van Bavel, Jay J., and Andrea Pereira. 2018. "The partisan brain: An Identity-based model of political belief." *Trends in Cognitive Sciences* 22 (3): 213-22.
Wu, Yuanyuan, Eric W. Ngai, Pengkun Wu, and Chong Wu. 2022. "Fake news on the internet: a literature review, synthesis and directions for future research." *Internet Research* 32 (5): 1662-1699.

Contemporary Debates in Political Philosophy

DOI 10.17473/2240-7987-2024-2-14

Global Poverty, Structural Change, and Role-Ideals

Olga Lenczewska[*] and Kate Yuan[**] [***]

Abstract

It has often been argued that charitable donations are not a sufficient response to global poverty; individuals need to address structural injustice. Proponents of the Effective Altruism (EA) movement have raised two main problems with this focus on structural injustice. In this paper, we respond to these concerns. The first problem raised by EA proponents is that focusing on structural injustice absolves individuals of any responsibility other than political ones. In response, we argue that discharging this duty requires more commitment than EA defenders think, and we do so by framing individual responsibility in global structural injustice through the lens of Robin Zheng's Role-Ideal Model (RIM). The second response given by EA proponents is that a focus on structural injustice does not provide concrete ways for any given individual to discharge such duties. To address this worry, we argue that RIM can be complemented with the Rawlsian account of moral maturation. This new framework makes it clear how individuals can form the right concept of justice and become responsible citizens who act in accordance with RIM.

Summary: Introduction. – I. Rethinking Responses to Global Poverty: Beyond Charitable Donations. – II. Effective Altruism's Objections to the Structural Change Approach. – II.I. The Undercommitment Objection. – II.II. The Intangibility Objection. – III. Moral Maturation and the Role-Ideal: Integrating Rawls. – III.I. The Rawlsian Framework to Develop into Role-Ideals. – III.II. The Justification for Extending the Rawlsian Framework onto the Global Scale. – Conclusion. – References.

[*] ORCID: 0000-0002-7977-8151.

[**] ORCID: 0009-0004-0923-7968.

[***] The authors contributed equally to this work; names are listed in a randomized order. The authors would like to thank Robert Smithson, Robin Zheng, and Adam Zweber, as well as the attendees of the conferences organized by the Long Island Philosophical Society, the North Carolina Philosophical Society, the South Carolina Philosophical Society, and the PPE Society, as well as two anonymous reviewers of this journal, for their valuable feedback on earlier drafts of this paper.

Introduction

In recent years, the issue of global poverty has come under increasing scrutiny. The world's affluent populations, who enjoy a significantly higher standard of living, have been criticized for their inability or unwillingness to make a meaningful impact on the lives of the world's poor. Effective Altruism (EA), a movement motivated by Peter Singer's (2009) positive duty account, aims to rectify this situation and has been one of the most prominent movements in the space of global justice. However, it has failed to consider affluent persons' negative duties towards the world's poor, namely the fact that we contribute to an unjust distributive structure that benefits us but does not provide the poor with basic necessities.

This paper contributes to the existing literature on this topic in the following ways. To begin with, we argue that common-sense philanthropy championed by the EA movement is not a sufficient response to global poverty. Affluent individuals should also understand and discharge their negative duty not to perpetuate structural injustice towards the world's poor (Pogge 2002). For the purpose of our paper, we define structural change broadly as any attempt that enduringly changes the distributional profile among a set of actors and can be large or small in scale, major or minor in impact. By understanding unjust structures as transcending formal organizations and as encompassing also the informal yet patterned sociocultural landscape, a more comprehensive assessment of distributive justice is possible.

Next, we address two responses given by defenders of the EA movement. The first is that focusing on structural injustice reduces individual responsibility to solely political actions. In contrast, we argue that fulfilling this duty involves a far greater commitment than EA defenders typically acknowledge. We frame this individual responsibility for addressing global structural injustice through Robin Zheng's Role-Ideal Model (RIM) (2018), which proposes that well-off individuals can work toward the ideal version of their social roles to effect structural change. With certain modifications, RIM highlights promising opportunities for motivated individuals to initiate structural change, even if they are the sole actors pushing for the change.

The second response given by EA defenders is that a focus on structural injustice does not provide concrete ways for any given individual to discharge such. To address this worry, we leverage Rawls' (1999a) account of moral maturation in order to suggest specific practical ways in which these duties can be discharged. Drawing on Rawls also allows us to suggest that well-off individuals should not only recognize their negative duties towards the

world's poor, but also be motivated to discharge such duties. This motivation may come from appropriately oriented social and educational institutions which provide the citizens with the right path of moral maturation that would involve the recognition of, and motivation to act on, their negative duties towards the world's poor. This new framework makes it clear how individuals can form the right concept of justice and become responsible citizens who develop into their role ideals and act in accordance with RIM. It also addresses a core limitation of RIM that the model "does not itself adjudicate between competing role-ideals" and thus does not explain how individuals can form the right concept of justice (Zheng 2018, 883). Supplementing RIM with an account of moral maturation helps us see how individuals can decide what to do: which roles to take up and how to fulfill this role.

I. Rethinking Responses to Global Poverty: Beyond Charitable Donations

Typically, philanthropic activities by well-off individuals are met with appreciation rather than doubt or criticism from society. Additionally, such activities are accompanied by tax benefits (Madoff 2016, 179-81). But when we examine the total distribution of charitable dollars, a pattern emerges that is hard to reconcile with redistributive outcomes (Reich 2018, 85). A recent study shows that at most one-third of charity is directed to providing for the needs of the poor. Among all the charitable giving, moreover, only 5% goes to the global poor (ibid., 88-9).

One of the most influential philosophical arguments for the affluent to donate to the poor globally comes from Peter Singer. Singer anchors our responsibility to the poor in a positive duty, namely that we should prevent the vast majority of people from living in such life-threatening poverty without incurring significant costs to ourselves (1972). Singer's EA movement has successfully leveraged the growing momentum of individual contribution to address global poverty. Between 2015 and 2021, around $420 million was donated each year, growing at 21% per year (Todd 2021). More strikingly, an estimated $30 billion of future donation was committed to EA causes, with a growth rate of about 20% each year (id. 2022).

Despite its success, EA has encountered a multitude of criticisms, particularly following November 2022 when a major donor to this movement misused billions of client funds (Mack 2022). The effective altruists' emphasis on 'earn-to-give' accepts capital accumulation as ethically unproblematic if do-

nated, absent forceful critique of the legitimacy of particular industries or money-making means (Conroy 2022). [1] More precisely, those who earn to give rely for their ability to give significant amounts of money to effective aid organizations on their privileged position within an unjust global economic order (Mills 2012, 5).

This phenomenon can be examined through Thomas Pogge's framework, which argues that affluent individuals contribute to an unjust distributive structure that privileges them at the expense of the global poor (2002). [2] Pogge outlines two duties for affluent individuals within his normative framework: (1) the duty to avoid further contributing to unjust structures, which he presents as achievable by promoting structural change, and (2) the duty to compensate for harms caused by redistributing unjustly gained benefits. Affluent individuals should not think that by making charitable donations they are practicing praiseworthy beneficence when their wealth is a result of global distributional injustice. The appropriate understanding of philanthropy under these circumstances is that it should serve the reparative aims of redressing the background wrongs of the unjust structures that produced the unfair distribution of resources in the first place. If a just global structure were in order, the well-off donors would have less income and wealth, and the intended beneficiaries would have more.

While Pogge (2017) sometimes frames promoting structural change as an optional strategy to fulfill the first duty, our approach views promoting structural change as an essential component of this duty. Specifically, we argue that the duty to stop the perpetuation of injustice is inseparable from the active commitment to structural reform, as passive non-participation is insufficient to meaningfully address the depth of existing injustices. In this paper, we therefore focus exclusively on the first duty, but interpret it in a stronger, forward-looking sense which emphasizes an obligation to disrupt structural injustices rather than merely avoiding their reinforcement. [3] Our position goes beyond

[1] Singer has cautioned against longtermist views (2021).

[2] For our purposes, the unjust distributive structure refers not only to a kind of organizational structure like codified institutions, but also to norms and common practices. Pogge, in one of his latest publications, concurs with this understanding of such an expansive delimitation (2023, 7). As an example of unjust distributive structure, the imposition of trade protectionism is estimated to inflict an approximate annual detriment of $100 billion upon people from the poorest countries. Additionally, the outflow of illicit financial resources exacts an added annual toll of $25 billion (Pogge 2010, Kar 2011).

[3] Young argues that forward-looking political responsibility matters because it focuses on proactive change rather than assigning blame. Traditional, backward-looking models of re-

Pogge's by asserting that fulfilling one's moral obligations in the face of global injustice requires a proactive stance on systemic reform rather than simply avoiding contributions to harm. This interpretation is more comprehensive because it expands moral responsibility to encompass not just avoiding harm but actively pursuing structural justice as a fundamental aspect of fulfilling the first duty. It emphasizes the importance of both individual actions and their cumulative collective impact, demonstrating that meaningful change requires ongoing structural engagement rather than isolated or one-time acts of avoidance of harm. Additionally, it is dynamic in its approach, advocating for sustained efforts to address structural injustices. This includes being adaptable and responsive to the evolving nature of these structures, rather than treating non-participation as a sufficient or final solution.

By clarifying our stance in this way, we also aim to distinguish our argument from prior literature, which has extensively examined the backward-looking duty of compensation, and instead center our analysis on the necessity of pursuing a forward-looking approach to justice. [4] While the second duty of compensation is significant, it is more achievable within existing frameworks and, in practice, aligns with many effective altruist efforts that focus on alleviating symptoms of injustice rather than its structural sources.

In contrast with the negative duty account, Singer's positive duty-based arguments are problematic in two interlinked ways. First, they contribute to the common-sense sentiment that donors can decide where their money goes, since 'it is their money, after all.' GiveWell, part of the EA movement that ranks charities' cost-effectiveness, recommends for donors to "choose the top charity (or charities) you prefer" (Givewell n.d.). [5] But Pogge's negative duty account suggests that what individual donors prefer should not matter. The wealth possessed by affluent individuals cannot be considered truly theirs since the current distribution of property across the world is widely regarded as unjust. [6] If I violate your property rights and have a debt to you that you

sponsibility (liability models) often emphasize fault and punishment, looking at past actions to determine who should compensate or be sanctioned (2004, 378-80).

[4] See Corvino and Pirni (2021) and Yuan (forthcoming) for discussion on the second duty.

[5] Effective Altruism does not afford donors absolute autonomy regarding the allocation of their contributions; nonetheless, it enforces fewer constraints compared to Pogge's negative duty.

[6] Various social distribution theories, including libertarianism, liberal-egalitarians, Kantian conceptions of property rights, suggest the existing pattern of property distribution is significantly unjust (Cordelli 2016, 242-4).

rightfully deserve, I have a reparative duty to return what I owe you (Cordelli 2016: 244-6). In such cases, the person who owes the debt has limited or no discretion in determining how to fulfill their obligation or who should be the recipient of the payment. As Gabriel puts it, "The idea of 'doing good' is itself problematic because it encourages people to believe that assistance is a matter of personal discretion rather than a moral responsibility, making collective action less likely" (Gabriel 2017, 468). [7]

Second, EA has unjustifiably neglected issues related to structural change that could address the root causes of poverty. [8] People engaged in the EA movement have been said to "leave untouched the power structures that create and maintain systems of poverty" (Clough 2015). Instead, EA has focused its attention on encouraging individuals to direct resources to organizations that directly aid people living in poverty. But by "focusing only on how they can do the most good within existing political and economic institutions," effective altruists have thereby "neglect[ed] the good that could be done by reforming these institutions" (Dietz 2019, 106) and, as a result, are unlikely to "develop an accurate understanding of systemic sources of poverty or to put pressure on their governments to reform political institutions that exacerbate it" (Gabriel 2017, 468). Being concerned with shaping individual actions for the sake of maximizing 'the lives one can save' prevents effective altruists from seeing the bigger picture. While rescuing individuals may seem like the most effective solution, it can also lead to a short-sighted, piecemeal approach that jumps from one crisis to the next without addressing the root causes of the problems we face. To bring about significant and lasting progress, we need to look beyond individual actions and work towards institutional and structural change.

[7] The language of charity problematically perpetuates moral hierarchy between benefactors and beneficiaries, masking how the affluent gain from the unjust distributive structure that is harming the global poor while diminishing their agency (Darnton and Kirk 201, 90, Hattori 2003, 229-47). Psychological research even indicates monetary giving can increase individualism while weakening communal motivations, thus dampening altruistic dispositions (Vohs, Mead and Goode 2006, 1154-6).

[8] For an extensive discussion of this objection, see Berkely 2018, Broi 2019, Clough 2015, Dietz 2019, Gabriel 2017, Herzog 2016.

II. Effective Altruism's Objections to the Structural Change Approach

Structural change is a promising approach for promoting justice and equality, as it can generate long-lasting benefits and open up opportunities for further structural change. [9] But EA defenders' have two main objections against the structural change approach: the first is the *undercommitment objection*, which suggests that concentrating on structural injustice releases individuals from non-political duties; the second is the *intangibility objection*, which posits that such a focus lacks tangible avenues for any given individual to fulfil these duties.

In this section, we expand Zheng's Role-Ideal Model (RIM) to clarify and reinforce the idea that individuals bear responsibility for promoting structural change through their social roles, which requires significant commitment. While Zheng's model emphasizes the need for collective action to address systemic injustice, we argue that RIM also underscores the role of individual responsibility, even when one's actions may seem isolated or small-scale. Our interpretation of RIM emphasizes that each person holds a duty to strive toward the ideal version of their social roles in ways that resist perpetuating injustice and promote structural reform. This expanded view not only broadens the scope of RIM but also counters EA's objections to structural approaches, by showing that individuals can meaningfully contribute to structural change, motivate others, and fulfill their responsibilities, even when large-scale reform seems intangible. By linking individual and collective responsibilities within RIM, we demonstrate that both modest and widespread efforts are essential for advancing justice.

II.I. The Undercommitment Objection

Effective altruists argue that directing attention to institutional and structural injustice is a way "not to worry too much that we might be acting wrongly when we spend significant amounts of money pursuing projects and interests that we care about, at least so long as we engage in enough political activ-

[9] Many discuss the importance of structural change. See Beck 2020, Berkey 2021, Corvino & Pirni 2021, Eckersley 2016, French and Wettstein 2014, Hayward 2017, Goodin & Barry 2021, Gould 2009, Jenkins 2021, Lu 2011, 2017, 2018, McKeown 2021, Neuhäuser 2014, Nussbaum 2009, 2011, Powers & Faden 2019, Reiman 2012, Sangiovanni 2018, Sankaran 2021, Schwenkenbecher 2021, Young 2009, 2011, Ypi 2017, and Zheng 2018, 2019.

ity in support of the necessary institutional change (e.g. voting for the right candidates, attending rallies, organizing, and perhaps even contributing some money to relevant political efforts)" (Berkey 2018, 168-9). This objection from effective altruists, however, significantly underestimates the level of commitment needed for individuals to pursue all the political activities. In fact, promoting structural change requires even more commitment than discharging political responsibilities alone.

To address this, Zheng's RIM, unlike Iris Marion Young's (2011) well-known Social Connection Model (SCM), claims that individuals should also be held responsible for the actions they carry out in performing their social roles, in addition to their political responsibilities that contribute to unjust global institutions (Corvino and Pirni 2021, 140). RIM postulates that, since any well-off individual is complicit in and benefits from structural injustice, individuals are responsible for structural injustice. In particular, they can alleviate structural injustice through their social roles. They have a responsibility to perform their social roles in a way that does not contribute to the creation or maintenance of unjust structures, since "social structures are built up from micro-level interpersonal interactions which are continually negotiated, enacted, and reenacted" (Zheng 2018, 874). This means that individuals must be aware of the expectations associated with their roles and actively seek to fulfill them in a way that promotes justice. In Zheng's words, "we are, each of us, individually responsible for structural injustice through and in virtue of our social roles…it is everyone's job to fight injustice because it is already their job to perform their roles well. In other words, it is one's job not just to be a teacher, but to be a good teacher" (ibid., 873-8).

In this sense, RIM requires significantly stronger commitments on individuals than effective altruists suggest, namely, that directing attention to structural injustice is an excuse not to divert resources from their personal projects but only to engage in enough political activity. Instead, RIM suggests that focusing on structural injustice requires a more formidable and all-encompassing commitment of one's life than the one posited by effective altruists.

II.II. The Intangibility Objection

EA defenders rightly observe that many moral and political philosophers emphasize the importance of addressing structural injustice but often present "moderate accounts of what individuals are obligated to do in response to the overwhelming injustice and suffering that continues to plague our world"

(Berkey 2018, 171). A key difficulty in defining individual obligations in this context lies in the belief that it is nearly impossible to determine how a person's actions directly contribute to systemic harms. The main contention is that tracing how limited actions interact with entrenched structural injustices is epistemically challenging, making it difficult to ground specific moral responsibilities and leaving ambiguity about what any individual should do. However, we argue that such epistemic uncertainty does not relieve affluent individuals of their duty to oppose structural injustice.

RIM offers an alternative by framing individual responsibility not in terms of specific causality or blame but rather as a commitment to fulfilling one's social roles in alignment with justice, even when direct causal links are unclear. RIM suggests that individuals, despite the constraints of their roles, still bear a responsibility to act in ways that promote justice. This model allows multiple individuals to share responsibility for a given instance of structural injustice, even if no single individual can be solely attributed as its cause. To fulfill these duties, individuals can strive to embody the ideal version of their social roles, such as acting as conscious consumers, informed voters, and responsible employers. As McKeown (2021) notes, "[e]ach role has a set of expectations about what the person will do in that role and normative beliefs about how they should act and be" (9). Thus, RIM responds to the intangibility objection by addressing a core limitation of Young's SCM, which has been criticized for failing to guide individuals on how to address structural injustice (Barry & Macdonald 2016; Hahn 2009; McKeown 2021).

However, Zheng emphasizes that "structural transformation is made possible when *all* individuals throughout the entire system push the boundaries of their social roles" and that "pressure must be applied throughout the *entire* system" (2018, 877; emphasis added). This raises a crucial question: What happens if no one but me engages in pushing these boundaries? Would my individual efforts be rendered futile in such a scenario? While it could be argued that everyone has a responsibility to challenge their role-boundaries through their role-ideal, the practical reality suggests that a significant portion, if not the majority, are unlikely to take up this challenge. As such, the prospect of any single individual effecting meaningful change remains intangible.

In light of this, we propose a refinement: even if only a few individuals (or even only one individual) were to exert such pressure, this still warrants a claim of structural change, albeit on a very modest scale. [10] Building on this modifica-

[10] For a related argument that people should take a stand against structural injustice even if it is likely to prove futile, see Goodin & Barry 2021.

tion, our interpretation of RIM now emphasizes three interlinked claims. First, the fact that others may not fulfill their duties does not lessen my own responsibility to act against structural injustice; each person bears an individual duty to avoid perpetuating injustice to the extent they can. Second, this duty to promote structural change includes joining and motivating others, as individual actions are where collective effort starts. [11] As such, my responsibility to push the boundaries of my role-ideals might also include motivating and inspiring others to do so as well. This is the kind of responsibility that I am able to discharge on my own, and my effort can have a small but incremental effect on structural change. This collective dimension acknowledges that while isolated actions may not create large-scale reform, individuals can inspire broader participation.

Third, even if one's efforts result in only modest or small-scale changes, this still fulfills an essential moral duty, as any movement toward justice—even incremental—is meaningful. The overwhelming difficulty to initiate any structural change by any individual could also come from the narrow conception which considers structural change as referring to significant changes to long-standing global policies, systems, and institutions that are deeply codified and entrenched in our various societies. But broadly defined, structural change refers to modifications made to the environment of a set of actors, which leads to a lasting impact on the distribution of power and opportunities among them. The nature of structural change can vary greatly, ranging from small-scale changes in a village to large-scale changes across the world, from social and cultural norms to legal regulations and institutional arrangements. [12] People mistakenly believe that only large and powerful groups, such as government officials, policymakers, and civil society leaders, can affect global structural changes. In light of our interdependence within the extensive global economic frameworks and the rapid progress of technological innovations that connect all of us, there are smaller-scale efforts any individual can do on their own to bring about structural change with global impact.

[11] Following Young, Zheng takes it to be a responsibility to join others in the collective effort of bringing about change. Zheng expressed this belief in personal correspondence with the authors.

[12] Given their respective definitions of structural injustice, Pogge (2023, 7) and Zheng (2018, 869-70) would agree structural change can be as small as parents choosing where to enroll their children. Zheng provides an example where one parent decides to enroll their child into a segregated school to combat racism and segregation because "even a handful of middle-class families made it less likely that a school would be neglected" (2018, 881).

To illustrate, consider a responsible employer who promotes structural change by trying her best to treat her employees fairly and transparently. While her actions may not overturn the inherent structural power asymmetries between employers and employees, she sets a role ideal and influence others in her sphere, contributing to gradual shifts in practices and norms. Such examples demonstrate that individuals can be held accountable for promoting structural change through their social roles, even if the direct impact is limited. This expanded conception of RIM thus allows for both collective and individual accountability, emphasizing that individuals have a responsibility to engage in structural reform at any scale, thereby countering EA's intangibility objection. In the following section, we will present practical examples of how individuals can engage in structural change efforts within their everyday roles.

III. Moral Maturation and the Role-Ideal: Integrating Rawls

III.I. The Rawlsian Framework to Develop into Role-Ideals

Once we recognize RIM to be a useful way to think about individual responsibility for structural injustice, we need a framework for understanding how we can become the kind of citizens who act in accordance with our role-ideals. We all occupy different social roles in our family, community, and society, and we are participants of various institutions, practices, and social orders. These roles have far-reaching global impacts—especially given our interconnectedness through large economic systems and fast-moving technological advances. How can people become individuals who feel responsible to strive for role ideals in order to promote structural justice on a global level and are motivated to act in ways that promote it?

To provide a framework that addresses this question, we turn to John Rawls's theory of moral development from *A Theory of Justice*, Part III. [13] Our reasons for appealing to Rawls' theory in particular are twofold. First, he draws heavily on the scientific-psychological work of Jean Piaget and Lawrence Kohlberg. Piaget and Kohlberg are credited with introducing the topic of moral development into psychology and their account of moral development

[13] Rawls's political philosophy largely influenced Pogge's own work (Pogge studied under Rawls and has dedicated some of this work to Rawls's theory).

remains the dominant paradigm in the field [14]—even though, of course, it has been challenged and complemented by various other models. Second, Rawls' three stages of moral development (as we will shortly see) center on one's relations to the progressively larger, or more encompassing, community that one shares and creates with others. It therefore suits RIM's focus on acting jointly with others and on the roles we play in our communal life—*social* roles. Rawls shows how one can become the sort of person who sees and treats others a moral equals. It is this egalitarian nature of his account that appeals to us. What Rawls claims in Part III of *A Theory of Justice* can help us see how the right kind of moral psychology can lead people to be motivated to promote structural change through their social roles. [15]

Rawls's conception of justice presented in *Theory* assumes a conception of the person as someone possessing two "moral powers": (1) the capacity to develop and pursue a conception of the good or happiness, and (2) the capacity to acquire a sense of justice. In Part I, Rawls proceeds on the assumption that adult citizens in a well-ordered society [16] (WOS) already possess these two moral powers. But how do we come to have these powers? Are they something we are born with or something that we have to learn? Rawls claims that these two fundamental characteristics of persons begin as mere capacities which have to be developed or realized throughout one's early life.

According to what Rawls writes in Part III of *Theory*, one must go through certain stages of moral development if one is to become a good member of a WOS and contribute to the inherent stability of the conception of justice that

[14] As Blum (1994, 185) writes, "[One type of moral development theories] is concerned with the adult capacities in which morality can be grounded, their development, and the specific childhood capacities that are their developmental precursors. Some of the most prominent of these theories are of a neo-Kantian nature, such as those of Rawls (1971) and Kohlberg (1981, 1984)."

[15] No global-level duties to help with the distribution of wealth in Rawls's major work in international justice, *The Law of Peoples*. According to *The Law of Peoples*, we have many similar, but not exactly the same, principles of duties towards citizens of other countries vs. our own country. There, Rawls does not argue for a substantive principle of economic distribution, such as the Difference Principle known from his work on domestic justice (Wenar 2004). We address this issue in section 3.2.

[16] Rawls defines a well-ordered society as "a society in which (1) everyone accepts and knows that the others accept the same principles of justice [as fairness], and (2) the basic social institutions generally satisfy and are generally known to satisfy these principles." (Rawls 1971/1999a, 4) Not all liberal democratic states qualify as well-ordered, but they can be considered "decent societies," which come close to WOS but do not satisfy all standards of Rawls' theory of justice as fairness. (Rawls 1999b)

governs the WOS. [17] In what follows, we elaborate on these stages and explain how they can contribute to acting in accordance with RIM.

For Rawls, the morality of authority is the first and most primitive phase of moral development, which a child born into a society should undergo. This stage is primarily characterized by the child's following concrete rules or commands issued by another person —an older child or an adult—which from the child's perspective are arbitrary and not connected with her own desires. In addition, Rawls explains the obedience of the child in the first stage by appealing to the emotional and affective bond between the child and the adult issuing rules or commands. [18] This authoritative person is someone whom the child trusts because of this person's affective and caring attitude toward the child (Rawls 1999a, 408). Thus we receive two explanations of obedience in this stage: the desire for avoiding punishment and receiving rewards and the willingness to follow those whom one trusts and from whom one receives affection. Importantly, a young child in this stage of moral development does not possess the notion of justice. Rawls characterizes the child in this phase as someone who "cannot comprehend the larger scheme of right and justice within which the rules addressed to him are justified" (Rawls 1999a, 408). Consequently, the child in this phase only possesses the moral characteristics of obedience, humility, and fidelity to authority (Rawls 1999a, 405-8).

Assuming we want to act in accordance with our role-ideals, what can we do during the Rawlsian first stage of ideal moral development in order to fos-

[17] Rawls's theory of moral development is part of a larger project Rawls pursues in Part III of *Theory*—this project is to show that a society regulated by his conception of justice, justice as fairness, is inherently stable. For a society to be stable means for it to be in a condition of equilibrium and able to return to such an equilibrium if some disruption takes place. According to Weithman (2010, 55, 102), unlike the kind of (imposed) stability that Plato or Hobbes were concerned with, which relies on heteronomous incentives and a sovereign who enforces obedience, Rawls is interested in stability that arises inherently from the desires and motivations of the citizens—desires and motivations shaped by the institutional setting of a WOS. This stability is the kind of stability in which—in Rawls's own description of a WOS—"inevitable deviations from justice are effectively corrected or held within tolerable bounds by forces within the system" (that is, within the citizens themselves) (Rawls 1971/1999a, 401). Among these forces, Rawls adds, "the sense of justice shared by the members of the community has a fundamental role" (Rawls 1971/1999a, 401; see also Forrester's (2019, 1-39) discussion on Rawls's analogy of society's stabilizing abilities to a game).

[18] Blum (1994, 196) characterizes the mechanism of this emotional and affective bond as "responsiveness" and emphasizes its importance for the subsequent developmental stages: "responsiveness in children is one developmental forerunner of the adult moral virtues of compassion, kindness, helpfulness, sympathy, and the like, in that these altruistic virtues as well as responsiveness involve altruistic motivation and sentiment toward others."

ter responsibility for promoting structural justice on a global level? Those of us whose social roles include 'parent' or 'guardian' of young children may strive to become ideal versions of this social role and guide the development of our children in a way that will promote structural justice in a variety of possible ways. [19] For example, we can tell our children that they already form a part of the global community and model globally ethical behaviors for them. We can choose to send our children to schools with greater diversity, encompassing broader international demographics. Such behaviors may include a genuine concern for global justice in the form of conversations about these topics or demonstrated donations to effective charities or to charities that have a tangible significance to the child. One instance of failing to fulfill one's role-ideal as a parent would be to teach children that they should distance themselves from children with a different skin color Hoffman (1976, 135). inquires about the emergence of children's genuine empathy for each other once they can differentiate between their own sense of self and that of others. He responds by suggesting that they realize the commonalities they share with others are more significant than the disparities. This perception of likeness forms the basis for a child's capacity to empathize with others. Even disparities in skin color do not inherently create a divide between children, except in situations where the child has been taught (for instance by parents) that skin color implies varying worth or demands separation.

The second stage of moral development, which Rawls calls the morality of association, consists in the gradual acquisition of skills required for social cooperation and for choosing among several heteronomous rules which sometimes come into conflict. This stage involves learning to see things from other people's perspectives and conducting oneself with reference to these perspectives, which makes cooperation possible. It also involves navigating social situations and commitments within a larger group of people. In this stage, the individual develops and cultivates the desire for forming ties of friendship with other members of her associations, and comes to obey the rules of the associations out of these ties of friendship. Rawls also characterizes this phase as the one when the child acquires the skills necessary for (however crude) social cooperation and for feeling mutual respect. Acquiring the morality of associa-

[19] Here, an objection could be raised: how can parents who were not raised in accordance with Rawls' model act as ideal versions of the social role as parents? Even if they are not capable of fulfilling this ideal, they might strive to become the best versions of themselves as parents as much as they can. If this progresses in a linear way, every next generation will approximate their ideal roles as parents more and more.

tion thus consists in recognizing that different points of view or perspectives exist in the minds of different people. These, in turn, lead people to have different desires, plans, and motives. Morality of association involves moral obedience for the sake of maintaining good social relations and receiving the approval of others in the association, as well as obeying any democratically accepted legal norms because they are socially useful (Rawls 1999a, 409-13).

Looking at the second stage of ideal moral development through the lens of RIM, it presents us with an opportunity to expand what we conceive of as our immediate social circle and the community in which we are embedded. This stage overlaps with late childhood, teenage years, and perhaps even early adulthood, and thus with the years during which an individual receives different levels of schooling. In order to fulfil one's ideal social role of 'student' in a way that fosters responsibility for global injustices, one can actively seek out communities outside one's own country. According to Blum, "[t]o be concerned for a friend, or for a community with which one closely identifies and of which one is a member, is to reach out…to what shares a part of one's own self and is implicated in one's sense of one's own identity" (1994, 195). This can be done by learning about other cultures (at school or on one's own), by forming one's identity as a global citizen who feels concerned about other members of the global community, and by exposing ourselves to people who differ from us. As Darwall writes, "it is impossible for individuals in racialized groups to relate to one another as equals and be mutually accountable for doing so, unless they encounter one another in daily life—in their neighborhoods and parks and other public spaces. The only way to abolish racial hierarchy and eliminate 'badges of slavery' is to establish relational equality, and that will require the abolition of racially segregated spaces" (forthcoming, 159-60). Hence, an ideal global citizen is someone who is aware that we all indirectly interact with many people around the world (through participating in economic activities, policy selection, media presence, and culture creation) and who engages in these activities in a way that treats these people in a just way. Some practical examples include traveling abroad or engaging in international virtual communities and expanding one's social ties to people from other countries. For instance, a college student in Europe can participate in the Erasmus exchange program, during which students live and study in another country for a semester or undertake an internship abroad. Other programs that promote global citizenship include international service learning trips, where students volunteer abroad and reflect on global inequities. Religious and secular organizations also sponsor international youth exchanges to build cross-cultural understanding. Virtual platforms, social media networks, and online

discussion forums have democratized the process of forming global connections, enabling individuals to engage in meaningful dialogues with people from diverse backgrounds without the confines of geographical borders. These digital interfaces provide a medium for sharing experiences, perspectives, and concerns, fostering a collective sense of global citizenship transcending physical limitations. As individuals partake in these transformative experiences, they are poised to develop a heightened awareness of the shared challenges and aspirations that unite humanity, underscoring the significance of collaborative efforts to address global issues collectively.

The final stage of Rawls's account of moral development, the morality of principles, is aspirational (regulative or ideal): adults may oscillate between the second and third stages for their entire life or approximate the third stage as they mature as adult citizens. This stage differs from the previous two: persons in this phase of moral development understand why they ought to act in accordance with principles of morality or justice and choose to do so without the need of any external coercion or incentives—autonomously. Persons in this phase of moral development understand that acting in accordance with principles of morality or justice involves acquiring a sense of the self as a member of a harmonious community of people who regard one another as moral persons (Rawls 1999a, 414-9).

The final stage of moral development presents an ideal model of how to look at others in this world no matter what one's exact social role (as an employee, parent, citizen, etc.) is because it involves generalizing moral rules into principles that can apply to every human being in the world, not just to people one is directly associated or familiar with. Kohlberg (1981) expresses a similar sentiment that a moral principle is a procedure or method called "ideal role taking" for making moral decisions. For instance, in ideal role taking, an agent figures out the right course of action by envisioning themselves in the place of everyone impacted by their potential decision (Blum 1994, 206). One of the moral principles we can learn and adopt, therefore, is to care about people globally and to help those who are in need regardless of where they live. In Moody-Adams's (2022, 4) terms, we can develop "human regard—a combination of compassionate concern and robust respect" to other global citizens. [20] As a citizen of a well-off country, one can perform such roles in a variety of ways. To return to the earlier example of the businessman: for those

[20] Moody-Adams (2022, 4) argues in a similar vein that for progressive social movements to happen, comprehending individual moral growth is prerequisite to embedding compassionate concern and robust regard for the disadvantaged within institutions and social practices.

who work at multinational corporations with operations in poor countries, their role as an ideal businessman ought to include considering if they are performing it in a globally just way, not just maximizing corporate interests. An ideal consumer, in turn, will be aware of being part of a global consumer chain and will avoid purchasing items made in unjust ways, such as in sweat shops. For others, it may involve looking out for the interests of their fellow global citizens, not just their compatriots. Ordinarily citizens can do so by caring and being knowledgeable about foreign policy, voting for politicians who want to address the struggles of the global poor, raising awareness of global issues in one's immediate community, or donating to the right charities.

III.II. The Justification for Extending the Rawlsian Framework onto the Global Scale

Before concluding this section, we want to address a number of objections to utilizing Rawls's moral-psychological account in the way we do, including the worry that Rawls himself did not wish to extend the scope of his framework of moral maturation onto global issues and the worry that his account of moral maturation might be inadequate to guide the development of liberal citizens because it is "comprehensive". We hope to show that even though Rawls himself did not believe that global justice requires of us any economic distributive principles such as the Difference Principle, his account of moral maturation from Part III of *A Theory of Justice* may arguably provide us with the tools needed to build such a framework, and that his account of moral maturation need not be seen as comprehensive in a way that precludes it from being used in a liberal society.

To begin with, Rawls's account of moral maturation might be seen as an inadequate framework to guide the development of liberal citizens because it is "comprehensive", i.e., relies on certain metaphysical and normative commitments that go beyond the scope of the "political" or of what public reason is supposed to determine. [21] Though Rawls never suggested this about moral maturation explicitly, [22] he did claim in his later work – in *Political Liberalism*

[21] Rawls distinguishes a comprehensive doctrines from a non-comprehensive political conception of justice grounded in public reason in "The Idea of Public Reason Revisited" (1997), reprinted in *Political Liberalism: Expanded Edition* (Rawls 2005: 440-90).

[22] Even though in the Introduction to *Political Liberalism* Rawls claimed that his account of stability from *Theory* III was comprehensive, he did not mention his account of moral matura-

– that certain elements of his earlier theory of justice from *Theory* were too reliant on Kant's moral philosophy and hence comprehensive, and he sought to remedy this in his later work. [23] Since the account of moral maturation does not explicitly appear in this later work, it might seem that it was part of what Rawls abandoned due to an unacceptably comprehensive nature. However, in what follows we will show that this is not the case by providing two arguments for viewing his account of moral maturation as non-comprehensive.

First, just because Rawls's account of moral maturation is Kantian, [24] it does not necessarily follow that it is comprehensive. A popular view in the Rawlsian literature is that by the time *Political Liberalism* was written, Rawls largely abandoned the Kantian components of his earlier view, as part of his attempt to present that view without relying on any particular comprehensive doctrine (Dreben 2002, Quong 2013, Rostbøll 2011, Wenar 1995). However, there is scholarly disagreement about whether various Kantian components were really abandoned and (relatedly) whether the presence of certain Kantian elements necessarily entails reliance on a particular comprehensive doctrine (Forst 2017, Taylor 2022). The latter is especially relevant to our argument. According to Rainer Forst (2017, 143), *Political Liberalism* "is best read as a Kantian view, that is, as one with conceptualizes a noncomprehensive, autonomous, moral grounded theory of political and social justice for a pluralistic society. It is noncomprehensive in that it neither rests on some metaphysical notion of human nature nor seeks to give guidance on questions of the good life." So just because Rawls's account of moral maturation is Kantian, it does not necessarily follow that it is comprehensive (especially since, and we will shortly show, Rawls endorsed this account in his later work). One reason for the view that Rawls's mature work can be Kantian without being comprehensive is that this work retains the Kantian notion of reasonability, both in continuing to describe citizens as "reasonable and rational" (2005, 450, 481, 487)

tion in particular. What is more, he endorsed this account later on – in his *Justice and Fairness: A Restatement* (Rawls 2001, 163).

[23] In "The Idea of Public Reason Revisited" (1997), reprinted in *Political Liberalism: Expanded Edition*, Rawls writes: "the content of public reason is given by a family of political conceptions of justice, and not by a single one. There are many liberalisms and related views, and therefore many forms of public reason specified by a family of reasonable political conceptions. Of these, justice as fairness, whatever its merits, is but one" (Rawls 2005, 450). In the Introduction to *Political Liberalism*, he also explicitly admits that *Theory* mistakenly relied on a comprehensive doctrine (Rawls 2005, xv).

[24] For an explanation of why Rawls's account of moral maturation is Kantian, see Lenczewska (forthcoming).

and in claiming that a non-comprehensive (freestanding and independently grounded) political conception of justice has the power normatively to determine which of the comprehensive doctrines are reasonable (acceptable), and which are not (Forst 2017, 128). This latter notion of reasonability is Kantian because, as Forst explains (2017, 128), it follows Kant "in emphasizing that both the categorical imperative and the principle of right had to be grounded completely independently of any doctrine of value leading to the good life (or *Glückseligkeit*) in order to take priority over them." And the former notion of reasonability (of citizens) is Kantian because it conceives of citizens as required to justify their reasons for organizing a basic structure in a particular way without appealing to their comprehensive doctrines, but on the basis of their common practical and public reason (Forst 2017, 129). Rawls's late work is also Kantian in retaining the moral conception of "full autonomy" in relation to constructing, through practical reason, political norms that no reasonable person could deny. This conception is not ethical in the sense of requiring metaphysical or value commitments that go beyond the scope of the political, but it is nonetheless moral because it is "connected to the grounds and normative quality of the political conception" (Forst 2017, 129).

Since various Kantian elements of Rawls's framework should not be seen as comprehensive just because they are in some ways Kantian, one might wonder if this is also the case with Rawls's (Kantian) account of moral development. We believe that this is indeed so. As we have seen in Section 3.1, developing one's moral capacities to its fullest by reaching the third and final state of moral development means, for Rawls, the ability to see others as reasonable and rational human beings who are free and equal. Crucially, one is able to extend this view of persons to *all* other citizens, not only to those with whom one has formed special ties of affection or association; this is what distinguishes this stage of moral development from the previous one. This ability to see others as reasonable and rational human beings who are free and equal, we believe, is compatible with the late-Rawlsian freestanding ideal of public reason, which encompasses a family of non-comprehensive, political conceptions of justice. This is because Rawls writes about these non-comprehensive, political conceptions that their "limiting feature (…) is the criterion of reciprocity, viewed as applied between *free and equal* citizens, themselves seen as *reasonable and rational*" (Rawls 2005, 450, emphasis added). In so doing, Rawls postulates that the kind of people who would be able to exercise public reason and participate in the procedure of political constructivism in order to determine (non-comprehensive) norms of justice are people who are reasonable and rational – and the very goal of moral maturation is to develop or real-

ize a person's two moral power: the capacity for a sense of justice (reasonability) and the capacity for forming a conception of the good compatible with a sense of justice (rationality) (Lenczewska forthcoming, §2).

The second reason why Rawls's account of moral maturation should not be viewed as comprehensive is that Rawls himself endorses this account in his later, mature work, in which he disavows any comprehensive elements of his previous works. Specifically, in *Justice as Fairness: A Restatement* he writes that "essential to the role of the family is the arrangement in a reasonable and effective way of the raising and caring for children, ensuring their moral development and education into the wider culture", and appends a footnote that refers to "*Theory*, §§70-76", i.e., to his account of moral maturation presented in *Theory* (2001, 162-3). In Section §59 of *Justice as Fairness*, titled "A Reasonable Moral Psychology", he also explicitly refers to and endorses his account of moral development from *Theory* III (2001, 196). By doing so on multiple occasions, he strongly suggests that his views on moral development of the citizens of WOS from his earlier work remain unchanged. And since he cares a great deal about the final form of his theory of justice to be political and non-comprehensive, he thereby also implies that his earlier account of moral maturation should not be seen as part of a comprehensive doctrine.

We will now move to another objection against using Rawls's account of moral maturation for our purposes in this paper, namely, to the claim that he himself does not wish to extend this account onto the global scale, but confines it to matters within particular nation-states. Arguably, extending Rawls's moral psychological account of moral maturation from *Theory* III onto a global scale goes against Rawls's own views regarding global justice from *The Law of Peoples*. After all, Rawls suggested in *The Law of Peoples* that there should be no equal distribution of goods at an international level. However, according to Pogge (1994, 195-7), Rawls made a mistake in his attempt to apply his theory of justice to the international sphere, and this mistake is relevant to our response to the above objection. Pogge argues that Rawls' theory should consider social and economic inequalities to be a criterion for global justice. Specifically, he could have adopted one of two following strategies: extending his two principles of justice to the international level, or starting with a "global original position," whereby parties in this original position would not know their nationality (Scraffe 2016, 207). Though Rawls did not pursue either strategy, believing that the institutional inequalities that exist at the state level are not equivalent to those at the international level, we here wish to suggest (with Pogge) that a consistent application of his theory from his work on domestic justice would commit him—and Rawlsians more gener-

ally—to extending the two principles of justice, and his theory of justice as fairness at large, to the international level. Ideally, citizens will grow up in a world that affirms the idea that we are so connected as one human community that we should pre-theoretically care about others even though they are far away and differ from us in profound ways. Nothing in Rawls's moral-psychological developmental framework precludes extending one's morality of principles onto a global scale.

Another objection to extending Rawls' moral maturation account onto the global scale would be that citizens of WOS cannot hope to receive the same as what they give to the poor—something that Rawls would see as necessary for reciprocal, egalitarian relationships. However, given the economically asymmetrical relationships between the affluent and the poor, we should not expect to receive from the poor what we give them. Citizens of well-off countries have more to share with those from poor countries, and more global means to do so, than is the case the other way around. Once the veil in a global original position is lifted, then, citizenship of more affluent nations would generate different (more stringent) global obligations than citizenship of poorer nations. Moreover, even if such reciprocal treatment is not received right away, the role of an extended Rawlsian global justice would be to gradually instill in these citizens an appropriate moral framework—especially once economic injustice is ameliorated. The moral treatment of individuals from far-away nations should not be contested simply due to prima facie, or initial, lack of reciprocity.

This is even more salient given the current economic, technological, and geographical interconnectedness humans face on a global scale. Given the increasingly global life citizens lead, Rawls's framework applied today would require a citizens to care about people who live far away, beyond one's national border. In the presence of the right kind of institutional (educational, familial, and socio-political) arrangements, one can grow up to see oneself as a member of a global, harmonious community of people who regard one another as moral persons. These members of the global community should respect one another's moral personality by treating others in the way that justice requires. Given what we argued for, Rawls's theory of moral maturation can and should be extended onto the global scale.

Furthermore, Rawls's account of moral maturation and, more specifically, his view on the ideal moral psychology of a developed citizen (who has attained the final stage of moral maturation) is compatible with Zheng's RIM in not guiding individuals as to what exactly they should do to fulfill their role-ideals and to push the boundary of their roles in the right direction. This is be-

cause, as we have argued above, the Rawlsian account does not adjudicate between various (reasonable) comprehensive doctrines, and hence it compatible with all of them. In so doing, it does not force individuals into particular role-ideals or into specific ways of fulfilling these role-ideals based, say, on specific comprehensive ethical, religious, or metaphysical commitments. This is an advantage of Zheng's framework and of Rawls's account: the individual pursuit of structural justice should be allowed to take on many forms and to stem from many motivations, so long as the general commitment to justice is retained – a commitment which is, of course, non-comprehensive in nature. While one might wonder whether the kind of role-ideals we have suggested in this paper inherently push individuals toward particular actions or whether they can accommodate diverse approaches to justice, we believe that the best approach is to allow them to do the latter – so long as these approaches are compatible with, and fall within, the broad Rawlsian framework (i.e., treat individuals as free, equal, reasonable, and rational in his sense of these terms). Role-ideals understood and defined this way will be able to accommodate the sort of pluralism that characterizes our world (although only of the reasonable kind). While this framework allows for potential disagreements arising among individuals about the best way to ideally fulfill their roles, we believe that in most cases there will simply be several ways of ideally fulfilling a particular role, not merely a single one.

We acknowledge that, in our current world, the institutional arrangements necessary to foster a robust global community are still developing. In the absence of a just global order, it may seem unrealistic or implausible to see oneself already as a member of an ideal global community. However, we believe that aspiring toward that ideal remains crucial. One can still strive to regard all persons, including globally distant strangers, as moral equals worthy of respect and consideration, even if existing institutions do not always reinforce those attitudes. The limitations of present social conditions do not negate the ethical claims that all of humanity has upon us. We must let the moral demand like that of RIM and the Rawlsian framework of moral maturation shape our attitudes and actions as best as possible under non-ideal circumstances.

Conclusion

As Rawls reminds us, justice requires that we think beyond our own self-interest and work toward a world where everyone has an equal opportunity to lead a fulfilling and dignified life. In this paper we have argued that common-

sense philanthropy championed by Effective Altruism is not a sufficient response to global poverty and, consequently, that well-off individuals should both recognize and be motivated to discharge their negative duty not to further contribute to the unjust distributive structure of our world. The Role-Ideal Model makes it evident that discharging this duty by promoting structural change is not as practically and epistemically difficult as it may seem. Though the model places more commitment on individual life other than political responsibilities, we have also shown how the Rawlsian framework for moral maturation can help us become ideal versions of the social roles we already occupy and identify with.

Though the present global order may fall short of the ideal, with concerted effort we can progressively reshape our conceptions of the social roles we occupy and reimagine our institutions from generation to generation, in order better to approximate an ideal of shared global community. If each generation dedicates itself to this task, slowly but surely social roles and practices will be brought into greater alignment with our moral duties to all people. This path may not be linear or smooth, but over time our social arrangements can be reformed to foster the global perspective necessary for justice. Despite current limitations, we must remain hopeful that our ideals as global citizens can gradually become the reality through intergenerational commitment to moral progress.

Works Cited

Barry, Christian, and Robert E. Goodin. 2021. "Responsibility for Structural Injustice: A Third Thought." *Politics, Philosophy and Economics* 20 (4): 339-356.

Barry, Christian, and Kate Macdonald. 2016. "How Should We Conceive of Individual Consumer Responsibility to Address Labour Injustices?" In *Global Justice and International Labour Rights*, edited by Yossi Dahan, Hanna Lerner, and Faina Milman-Sivan, 123-140. Cambridge: Cambridge University Press.

Beck, Valentin. 2020. "Two Forms of Responsibility: Reassessing Young on Structural Injustice." *Critical Review of International Social and Political Philosophy* 26 (6): 918-941.

Berkey, Brian. 2018. "The Institutional Critique of Effective Altruism." *Utilitas* 30 (2): 143-171.

Berkey, Brian. 2021. "Sweatshops, Structural Injustice, and the Wrong of Exploitation: Why Multinational Corporations Have Positive Duties to the Global Poor." *Journal of Business Ethics* 169 (1): 43-56.

Broi, Antonin. 2019. "Effective Altruism and Systemic Change." *Utilitas* 31 (3): 262-276.

Calder, Todd. 2010. "Shared Responsibility, Global Structural Injustice, and Restitution." *Social Theory and Practice* 36 (2): 263-290.
Clemens, Michael A. 2011. "Economics and Emigration: Trillion-Dollar Bills on the Sidewalk?" *The Journal of Economic Perspectives* 25 (3): 83-106.
Clough, Emily. 2015. "Effective Altruism's Political Blind Spot." *Boston Review.* https://www.bostonreview.net.
Conroy, J. Oliver. 2022. "Power-Hungry Robots, Space Colonization, Cyborgs: Inside the Bizarre World of 'Longtermism.'" *The Guardian*, November 20, 2022. https://www.theguardian.com.
Cordelli, Chiara. 2020. *The Privatized State.* Princeton, NJ: Princeton University Press.
Corvino, Fausto, and Alberto Pirni. 2021. "Discharging the Moral Responsibility for Collective Unjust Enrichment in the Global Economy." *Theoria: An International Journal for Theory, History and Foundations of Science* 36 (1): 139-158.
Darnton, Andrew, and Martin Kirk. 2011. *Finding Frames: New Ways to Engage the UK Public in Global Poverty.* London: Bond.
Darwall, Stephen. Forthcoming. *The Heart and Its Attitudes.*
Dietz, Alexander. 2019. "Effective Altruism and Collective Obligations." *Utilitas* 31 (1): 106-115.
Eckersley, Robin. 2016. "Responsibility for Climate Change as Structural Injustice." In *Oxford Handbook of Environmental Political Theory*, edited by Teena Gabrielson, Cheryl Hall, John M. Meyer, and David Schlosberg, 321-335. Oxford: Oxford University Press.
Forrester, Katrina. 2019. *In the Shadow of Justice.* Princeton, NJ: Princeton University Press.
French, Peter A., and Howard K. Wettstein, eds. 2014. *Forward-Looking Collective Responsibility.* Midwest Studies in Philosophy, Vol. 38. Oxford: Wiley-Blackwell.
Gabriel, Iason. 2016. "Effective Altruism and its Critics." *Journal of Applied Philosophy* 34 (4): 457-473.
GiveWell. n.d. Accessed August 24, 2023. https://www.givewell.org.
Giving USA. 2023. "The Annual Report on Philanthropy for the Year 2022."
Goffman, Erving. 1959. *The Presentation of Self in Everyday Life.* New York, NY: Anchor Books.
Gould, Carol C. 2009. "Varieties of Global Responsibility: Social Connection, Human Rights, and Transnational Solidarity." In *Dancing with Iris*, edited by Ann Ferguson and Mechthild Nagel, 199-212. Oxford: Oxford University Press.
Hannah-Jones, Nikole. 2016. "Choosing a School for My Daughter in a Segregated City." *New York Times.* https://www.nytimes.com.
Hahn, Henning. 2009. "The Global Consequence of Participatory Responsibility." *Journal of Global Ethics* 5 (1): 43-56.
Hattori, Tomohisa. 2003. "The Moral Politics of Foreign Aid." *Review of International Studies* 29 (2): 229-247.

Hayward, Clarissa R. 2017. “Responsibility and Ignorance: On Dismantling Structural Injustice.” *Journal of Politics* 79: 396-408.
Herzog, Lisa. 2016. “Can ‘Effective Altruism’ Really Change the World?” *Open Democracy.* https://www.opendemocracy.net.
Hoffman, Martin L. 1976. “Empathy, Role-Taking, Guilt and Development of Altruistic Motives.” In *Moral Development and Behavior*, edited by Thomas Lickona, 119-148. New York: Holt, Rinehart, and Winston.
Jenkins, David. 2020. “Understanding and Fighting Structural Injustice.” *Journal of Social Philosophy* 52 (4): 569-586.
Kar, Dev. 2011. “Illicit Financial Flows from the Least Developed Countries: 1990-2008.” *United Nations Development Programme Discussion Paper.*
Kohlberg, Lawrence. 1981. *Essays in Moral Development, Vol. 1: The Philosophy of Moral Development.* San Francisco: Harper and Row.
Lenczewska, Olga. Forthcoming 2026. “Developing Politically Stable Societies: Kant and Rawls on Moral Maturation.” *Social Theory and Practice* 52(1).
Lu, Catherine. 2011. “Colonialism as Structural Injustice: Historical Responsibility and Contemporary Redress.” *Journal of Political Philosophy* 19: 261-281.
Lu, Catherine. 2017. *Justice and Reconciliation in World Politics.* Cambridge: Cambridge University Press.
Lu, Catherine. 2018. “Responsibility, Structural Injustice, and Structural Transformation.” *Ethics and Global Politics* 11 (1): 42-57.
Mack, Eric. 2022. “The Fall of FTX and Sam Bankman-Fried: A Timeline.” *CNET.* https://www.cnet.com.
Madoff, Ray D. 2016. “When Is Philanthropy?: How the Tax Code’s Answer to This Question Has Given Rise to the Growth of Donor-Advised Funds and Why It’s a Problem.” In *Philanthropy in Democratic Societies: History, Institutions, Values*, edited by Rob Reich, Chiara Cordelli, and Lucy Bernholz, 176-197. Chicago, IL: University of Chicago Press.
McKeown, Maeve. 2018. “ETMP Discussion of Robin Zheng’s ‘What is My Role in Changing the System? A New Model of Responsibility for Structural Injustice.’” *PEA Soup.* https://peasoup.princeton.edu.
McKeown, Maeve. 2021. “Structural Injustice.” *Philosophy Compass* 16 (7): 1-14.
Mills, Pete. 2012. “The Ethical Careers Debate: A Discussion between Ben Todd, Sebastian Farquhar, and Pete Mills.” In *The Oxford Left Review*, edited by Tom Cutterham, 4-9.
Moody-Adams, Michele. 2022. *Making Space for Justice: Social Movements, Collective Imagination, and Political Hope.* New York, NY: Columbia University Press.
Neuhäuser, Christian. 2014. “Structural Injustice and the Distribution of Forward-Looking Responsibility.” *Midwest Studies in Philosophy* 38 (1): 232-251.
Pogge, Thomas. 1994. “An Egalitarian Law of Peoples.” *Philosophy & Public Affairs* 23 (3): 195-224.

Pogge, Thomas. 2002. *World Poverty and Human Rights – Cosmopolitan Responsibilities and Reforms.* Malden, MA: Blackwell Publishers Inc.
Pogge, Thomas. 2006. "Moral Priorities for International Human Rights NGOs." In *Ethics in Action*, edited by Daniel A. Bell and Jean-Marc Coicaud, 218-56. Cambridge: Cambridge University Press.
Pogge, Thomas. 2010. *Politics as Usual: What Lies Behind the Pro-Poor Rhetoric.* Cambridge: Polity Press.
Pogge, Thomas. 2014. "Are We Violating the Human Rights of the World's Poor?" *Yale Human Rights & Development Law Journal*: 40-72.
Pogge, Thomas. 2017. "Fighting Global Poverty." *International Journal of Law in Context* 13 (4): 512-526.
Pogge, Thomas. 2023. "Poverty." In *Encyclopedia of the Philosophy of Law and Social Philosophy*, edited by Mortimer Sellers and Stephan Kirste. Springer, Dordrecht. DOI: 10.1007/978-94-007-6730-0_1077-1.
Powers, Madison, and Ruth Faden. 2019. *Structural Injustice: Power, Advantage, and Human Rights.* New York, NY: Oxford University Press.
Rawls, John. 1999a. *A Theory of Justice: Revised Edition.* Cambridge, MA: Harvard University Press.
Rawls, John. 1999b. *The Law of Peoples.* Cambridge, MA: Harvard University Press.
Rawls, John. 2001. *Justice as Fairness: A Restatement.* Cambridge, MA: Harvard University Press.
Rawls, John. 2005. *Political Liberalism: Expanded Edition.* New York, NY: Columbia University Press.
Reich, Rob. 2018. *Just Giving: Why Philanthropy Is Failing Democracy and How It Can Do Better.* Princeton, NJ: Princeton University Press.
Reiman, Jefferey. 2012. "The Structure of Structural Injustice: Thoughts on Iris Marion Young's Responsibility for Justice." *Social Theory and Practice* 38: 738-751.
Sangiovanni, Andrea. 2018. "Structural Injustice and Individual Responsibility." *Journal of Social Philosophy* 49 (3): 461-483.
Sankaran, Kirun. 2021. "Structural Injustice and the Tyranny of Scales." *Journal of Moral Philosophy* 18 (5): 445-472.
Schwenkenbecher, Anne. 2021. "Structural Injustice and Massively Shared Obligations." *Journal of Applied Philosophy* 38 (1): 1-16.
Scraffe, Eric. 2016. "'A New Philosophy for International Law' and Dworkin's Political Realism." *Canadian Journal of Law and Jurisprudence* 29 (1): 191-213.
Singer, Peter. 1972. "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1 (3): 229-243.
Singer, Peter. 2009. *The Life You Can Save: Acting Now to End World Poverty.* New York: Random House.
Singer, Peter. 2021. "The Hinge of History." *Project Syndicate.* https://www.project-syndicate.org.

Todd, Benjamin. 2021. "Is Effective Altruism Growing? An Update on the Stock of Funding Vs People." *80,000 Hours.* https://80000hours.org.
Todd, Benjamin. 2022. "EA Financial Resources Are Down Maybe ~2x Since I Wrote This." *Twitter*, August 20, 2022. https://twitter.com.
Vohs, Kathleen, Nicole Mead, and Miranda Goode. 2006. "The Psychological Consequences of Money." *Science* 314 (5802): 1154-1156.
Weithman, Paul. 2010. *Why Political Liberalism? On John Rawls's Political Turn.* Oxford: Oxford University Press.
Wenar, Leif. 2004. "The Unity of Rawls's Work." *Journal of Moral Philosophy* 1 (3): 265-275.
Young, Iris M. 2004. "Responsibility and Global Labor Justice." *The Journal of Political Philosophy* 12 (4): 365–388.
Young, Iris M. 2009. "Structural Injustice and the Politics of Difference." In *Contemporary Debates in Political Philosophy*, edited by Thomas Christiano and John Christman, 362–383. Oxford: Wiley-Blackwell.
Ypi, Lea. 2017. "Structural Injustice and the Place of Attachment." *Journal of Practical Ethics* 5 (1): 1-21.
Yuan, Kate. Forthcoming. "Global Justice: From Institutional to Individual Principles." *Social Theory and Practice*.
Zheng, Robin. 2018. "What is My Role in Changing the System? A New Model of Responsibility for Structural Injustice." *Ethical Theory and Moral Practice* 21: 869-885.
Zheng, Robin. 2019. "What Kind of Responsibility Do We Have for Fighting Injustice? A Moral-Theoretic Perspective on the Social Connection Model." *Critical Horizons* 20: 109-126.

DOI 10.17473/2240-7987-2024-2-15

Strawsonian Responsibility: Three Critiques from the Margins

Michelle Ciurria*

Abstract

In a landmark philosophy paper (1963), P.F. Strawson argued that moral responsibility is a matter of being able to participate in a "moral community" structured by "reactive attitudes" such as resentment, gratitude, forgiveness, and hurt feelings. On this framework, many, but not all, people are members of the moral community. Exceptions include "the insane," "young children," and, by extension, other-than-human animals. To be sure, these exceptions capture a commonsense way of thinking about responsibility, at least for the average Western, educated, industrialized, rich, and democratic (WEIRD) citizen. But "commonsese" reflects a particular cultural situation, and for Strawson, this was the situation of a nondisabled, white, cisgender, male, tenured philosopher at Oxford University in the 1960s. As marginalized philosophers increasingly migrate into the field of responsibility theory, they are calling into question Strawson's exemptions and proposing transformative changes. In this paper, I draw on emerging interdisciplinary and intercultural critiques to challenge Strawson's exclusion of (1) neurodivergent disabled people, (2) other-than-human-animals, and (3) young children from the moral community. By extension, these critiques problematize the "capacities criterion," which states that responsibility is an exclusive or canonical property of neurotypical adult humans. The aim of this paper is to correct an epistemic injustice in the philosophical literature by elevating marginalized standpoints.

Summary: Introduction. – I. Neurodivergent Disabilities. – II. Other-than-human Animals. – III. Young Children. – IV. Ameliorative Responsibility. – Works Cited.

Introduction

In 1963, P. F. Strawson published one of the 20th Century's most influential philosophy papers on moral responsibility. This paper initiated a paradigm

* ORCID: 0000-0002-8722-1003.

shift in the way philosophers think about responsibility, redirecting the focus from metaphysics to social theory (Holroyd 2022). One of the reasons for the enduring popularity of Strawson's essay is its resonance with prevailing social practices. Strawson both explained and justified the attribution (or denial) of responsibility in social institutions raging from the prison to the education system to the family. Therefore, if Strawson's framework is flawed, the consequences are far-reaching.

Strawson's landmark essay maintains that holding someone morally responsible involves seeing the person as a proper target of the "reactive attitudes," such as gratitude, resentment, forgiveness, anger, and hurt feelings. While most people are susceptible to these attitudes, there are notable exceptions, including individuals with (in Strawson's words) "compulsions," "insanity," "less extreme forms of psychological disorder," and "young children." Extrapolating from these cases, we can infer that other-than-human animals, [1] like dogs, cats, and horses, would likewise be exempt from the reactive attitudes. The standard justification for these exemptions is that these three groups lack the capacity to participate in interpersonal relationships and thereby contribute to the "moral community." As Matthew Talbert explains in the *Standard Encyclopaedia of Philosophy*,

> For Strawson, the most important group of exempt agents includes those who are, at least for a time, significantly impaired for normal interpersonal relationships. These agents may be children, or psychologically impaired like the "schizophrenic"; they may exhibit "purely compulsive behaviour", or their minds may have "been systematically perverted" (Talbert 2020, citing Strawson 1963).

To be sure, Strawson's exemptions capture a commonsense way of thinking about who is and is not responsible, at least for the average Western, educated, industrialized, rich, and democratic (WEIRD) citizen (Heinrich et al. 2010). However, commonsense intuitions reflect a particular cultural standpoint, and those of tenured professors like Strawson are no exception. As Robin Dembroff has observed, the "pretheoretical concepts and terms" that prevail in academic philosophy are, in general, those of the "culturally powerful," not "the commonsense of the racialized, poor, queer, transgender, or disabled," which are "considered philosophically irrelevant 'ideology,' 'activism,' or

[1] I use the term "other-than-human" rather than "non-human" to emphasize that species other than humans should not be defined in relation to humans, but should be understood on their own terms.

'delusion'" by comparison (2020: 403). As more diverse philosophers migrate into the field of moral responsibility – the historical dominion of privileged, able-bodied, white men – they are using their own commonsense to challenge the intuitive plausibility of Strawson's exemptions.

In this paper, I will draw on recent intercultural and interdisciplinary critiques of Strawsonian responsibility to challenge Strawson's three exclusions from the moral community: those of (a) non-neurotypical people, (b) young children, and (c) other-than-human animals. Implicitly, these exemptions share a common foundation in the notion that moral responsibility requires capacities uniquely or canonically associated with neurotypical adult humans. Therefore, non-neurotypical, non-adult, and non-human beings are exempt from the domain of responsibility. Following Stephanie Jenkins, I will call this assumption "the capacities criterion" (2024: 378). The capacities criterion defines "animals and people with disabilities (especially…. cognitive disabilities)" as moral patients and dependents, giving rise to "an exclusionary ethic" that relies on binary distinctions and hierarchies (383). Strawson's exclusions – which are widely accepted in moral philosophy – hang on this philosophical principle, which give us good reason to re-evaluate them jointly, albeit from the critical standpoints of those most affected by them. The purpose of this critical re-assessment is to enhance "epistemic justice" or fairness in the distribution of credibility, trust, and respect (Fricker 2011). This, in turn, will foster a more democratic arena of debate.

A common theme amongst the emerging critiques of Strawson is that, rather than emphasizing the strongest points of Strawson's work – such as his reappraisal of interpersonal relationships, which had previously been overshadowed by, in his words, the "panicky metaphysics" of analytic philosophers – they address the worst parts of his work: namely, the idealization of liberal policies and the corresponding absence of critiques that would explain why our responsibility practices are structurally unjust. In other words, these critiques share an emphasis on what Charles Mills (2007) referred to as "non-ideal theory," a critical approach to asymmetries of power that allows us to analyse and address systemic harms. Whereas "ideal theory" searches for moral principles under ideal conditions, "non-ideal theory" addresses injustices in the real world by attending to oppressed people's testimony. This is not to say that we can learn nothing from Strawson, but we should not discount the epistemic value of critical perspectives.

One question that Strawson's critics try to answer is, how can we use moral responsibility as a source of resistance to dismantle "the master's house," the dominant frameworks, methodologies, and ideologies that maintain oppres-

sion (Lorde 1984)? How can we collectively develop, to use Sally Haslanger's term (2005), an "ameliorative" approach to responsibility – one that leverages the constitutive parts of the responsibility system (reactive attitudes, relationships, norms) to alleviate rather than enforce oppression? This paper synthesizes a variety of answers to this question, offering an intersectional analysis. Together, these critiques provide multiple challenges to the capacities criterion, paving the way for a more inclusive, intercultural, and relational framework.

Before proceeding, I should clarify that this is not intended as a comprehensive critique of Strawson's exemptions, much less of the capacities criterion, which has a much broader scope. My aims are more modest: to survey some of the most recent critiques of Strawson, bringing attention to an emerging agreement that the Strawsonian paradigm is not as common-sensical or self-evident as some may believe, and showing that the hegemony of this paradigm has stifled constructive debate about what oppressed groups can contribute to society.

I. Neurodivergent Disabilities

The first set of critiques challenges Strawson's exclusion of neurodivergent people from the moral community. In critical disability theory, neurodivergence is often understood as a form of disability because neurological differences attract ableist discrimination, making them disabilities on the social model (Tremain 2017, 2024). [2] However, not everyone agrees with this classification. For simplicity, I will restrict my focus to neurodivergent disabilities: neurological variations that make one susceptible to ableist discrimination, and thus qualify as (socially constructed) disabilities. This term encompasses neurological variations that would fall under Strawson's labels of "insane" and "mentally disordered." Philosophers working in the overlapping fields of critical disability theory, neurodiversity theory, and Mad studies have raised objections to Strawson's exclusion of this category, which I will overview here.

To begin, Jules Holroyd offers a helpful explanation of why Strawsonian "conventionalism" is ill-equipped to accommodate neurodivergent perspectives (2024, 20). (Conventionalists, unlike some revisionists, agree with

[2] In brief, the social model of disability defines disability as a result of social barriers and discrimination, whereas the medical model defines disability as an individual impairment, limitation, or pathology (Oliver 2013).

Strawson's endorsement of the capacities criterion). When Strawson shifted the debate from metaphysics to social theory, he proposed that moral responsibility should be understood as "practice-dependent," meaning that its definition should depend on the internal practices of a given culture or community (2024: 1). Thus, in cultures that practice corporal punishment, sentencing someone to death for first-degree murder would be admirable by virtue of the culture's general acceptance of that practice, regardless of what other cultures think of the death penalty. In Strawson's exact words, the practice of responsibility "neither calls for, not permits, an external 'rational' justification." Expanding on this proposal, Victoria McGeer (2019) compares being responsible to being fashionable. As Holroyd describes this analogy,

> There are facts, at any particular time or in any particular context, about what is fashionable. Something or someone can have the property of being fashionable. But these facts, or possession of the property, are wholly determined by the norms of fashion and by the practices involved in it. There is no independent, objective property that fashion-related practices track (Holroyd 2024, 13).

What it means to be fashionable, in effect, is decided by "the teams of fashionistas coordinating the Paris fashion week" (Holroyd 2024, 14). David Shoemaker echoes this view, affirming that "there simply is no question as to [the responsibility system's] correctness or incorrectness from an external standpoint" (2017, 482). This reflects the conventional Strawsonian wisdom that the definition of responsibility, and thus its proper extension and implementation, is essentially a matter of cultural consensus. Cultural insiders determine what it means to be responsible, and the standpoints of outsiders are deemed irrelevant.

Having said this, Strawsonians acknowledge that we *can* be wrong about what it means to be responsible, but whether we're wrong depends on whether our judgements track the cultural consensus. The practice-dependent view doesn't seem to allow that a culture's responsibility practices could be systematically erroneous or structurally unjust. On a related note, by discounting "external standpoints," the practice-dependent view seems to restrict social change to incremental reform within the system – that is, piecemeal adjustments consistent with the dominant liberal tradition in the Global North. Indeed, McGeer, R. J. Wallace (1994), Michael McKenna (2012), and other "moderates" concur that revisions to the cultural definition of responsibility should be, in Manual Vargas' terms, "conservative" (2004, 230) and "modestly revisionary" (228), not radical or revolutionary. The appeal to incremental

change within the existing social order – which Mills describes as a "domination contract" wherein cultural elites rule over cultural subordinates (2017, 36) – aligns with the liberal contractarian tradition that Strawson favoured: one denounced by the likes of Mills, Carole Pateman (1988), and Stacy Simplican (2016) as structurally racist, sexist, ableist, and altogether counter-revolutionary. This "liberal reformist" approach (Chapman 2024) clashes with the preferred political strategies of critical disability theorists, Indigenous feminists, Marxists, and other revolutionaries who call for nothing short of a popular uprising against the ruling classes. As Martin Luther King Jr. said, "justice delayed is justice denied": incremental reform with an unjust system merely prolongs injustice and is an injustice in itself. In contrast to the liberal contractarian tradition, marginalized cultures tend to endorse a more revolutionary approach – one that aims to disrupt and reconstruct social arrangements.

Another issue with practice-dependent theories is that, even if they recognise a need for modest and incremental "recalibrations" to the system, to use McGeer's language (2019), they don't explain how we are supposed to arbitrate between conflicting cultural practices. Should we see lakes and streams as responsible agents, as Indigenous environmental feminists do, or is this perspective too "external," and therefore irrelevant, to the Western paradigm? How should we even approach this dispute? If we defer to the culture consensus of the Imperial West (to which Strawson belonged), then Indigenous feminisms are dismissed as, in Debroff's terms, mere "ideology," "activism," or "delusion," outside of the "space of reasons." Holroyd argues that Strawsonians must bridge this epistemic gap by providing "tools for articulating and addressing… oppressive norms, structures, and institutions" in a culture's responsibility system (2024, 20). The proposed revision includes a mandate to engage with marginalised cultures and address asymmetries of power between different communities and cultural frameworks, fostering epistemic justice and constructive friction.

Since Strawson's view does not include such a mandate – indeed, it seems to license epistemic injustice by discounting external standpoints – it can be seen as a form of ideal theory, a liberal contractarian approach that "abstracts away from actual injustices, and so lacks the resources to understand, much less address or provide remedy for, real world injustices" (Holroyd 2024, 20). Ideal theorists, whether intentionally or not, reproduce structural injustices by failing to acknowledge their existence. In Mills' terms, they misrepresent the "social ontology" as relatively fair, equitable, and governed by good-faith policymakers, thereby making it impossible to confront structural injustices (2017, xiv).

When philosophers discount the "external standpoints" of neurodivergent disabled people, they risk reinforcing ableist assumptions and double-binds. Holroyd observes that disabled people tend to be reduced to one of two stereotypes: "supercrips," who are celebrated for heroically overcoming their impairments, or "sadcrips," who are pitied for tragically succumbing to them. The first stereotype evokes patronizing praise, while the second elicits condescending pity. Either way, disabled people are denied the dignity of being seen as complex human beings whose lives cannot be reduced to "flat stories" of heroism or misery (Lackey 2024). Modest revisions to the responsibility system are insufficient to dismantle these "double-binds," which Marylin Frye (1983) defines as situations in which all available choices lead to oppression. Double-binds are a structural feature of the responsibility system, not a bug. As such, incremental reform within the system will not eliminate them. A more radical paradigm shift – toward a politics of liberation – is needed.

While Strawson is silent on ableist oppression, he writes extensively about neurodivergent disability, which he treats as an impairment and limit on moral agency. Accordingly, Shelley Tremain notes that, "in Strawson…, the parameters of the moral domain are delineated and secured through the exclusion of disabled people" (2014, 5). Strawson's stance on neurodivergent disability, in fact, replicates the *de facto* exclusion of disabled people from status-conferring institutions ranging from the family to the workforce to the education system. Disabled people, for instance, are "much less likely to be employed than those with no disability" (Bureau of Labor Statistic 2020), but much more likely to be incarcerated and involuntarily hospitalised (Prison Policy Initiative 2022), and thus removed from their communities. Autistic people are a representative example, with an estimated 25-30% facing unemployment, and about 12% being institutionalized (Farley et al. 2019). This is despite the fact that disabled people have a legal right to live in their communities with appropriate support, and research shows that at-home care costs less than institutionalization. Nonetheless, disabled people continue to be "medically kidnapped" due to "institutional bias" in favor of expensive modes of disabled segregation (Mizner et al. 2021, writing for the ACLU). Thus, neurodivergent disabled people are not, as Strawson imagined, exempt from their communities, but are instead forcibly *exempted* by ableist attitudes and policies. Strawson's analysis misconstrues these forms of ableist exclusion as natural, apolitical phenomena rather than structural injustices.

Strawson's embrace of the medical model has practical implications. In particular, it supports his contention that "insane" and "mentally disordered" people are proper objects of "treatment," "management," "training," and "so-

cial policy," as well as risks to be "avoided" and "taken precautionary account of," as opposed to inclusive members of the moral community. In other words, neurodivergent disabled people *ought* to be excluded from the moral community due to their tragic impairments. Their lives, furthermore, ought to be dictated by a social contract designed by neurotypical/nondisabled cultural insiders – that is, fully responsible and rational adults – and this contract should provide some form of medical or legal intervention, not political solidarity or disability justice. This reductive medical perspective on neurodivergent disability matches the *de facto* cultural consensus that disability is an impairment that prevents people from being productive members of society.

The medical model of disability was especially prevalent in Strawson's time, when lobotomies, forced sterilisation, ice baths, and other forms of torture were considered humane treatments for "lunatics" and "neurotics," who were seen as contributing nothing of value to society. The "social policies" of the 1960s that Strawson recommended were, in hindsight, eugenic policies that most modern readers would find appallingly ableist. These policies, furthermore, intersected with cultural notions of Blackness, queerness, and poverty as "disabling conditions," making these other groups susceptible to the same kinds of eugenic violence as "the insane" (e.g., overincarceration, institutionalization, unemployment). Indeed, it is impossible to disentangle ableist eugenics from other intersections of oppression, since ableism is and has always been weaponized against other marginalized populations to control, discipline, and exploit them. By uncritically recommending the social policies of the time, Strawson lent credibility to eugenic practices that entrenched multiple injustices.

August Gorman agrees that Strawsonian conventionalism perpetuates neuronormative/ableist stereotypes by failing to engage with the neurodiversity model, which rejects the medicalization and individualization of neurodivergent disability. Gorman specifically critiques the Strawsonian assumption that non-neurotypical people are proper objects of sympathy and pity: "This dehumanizing pity is on full display in P.F. Strawson's view, which is arguably the most widely adopted framework for thinking about the nature of moral responsibility" (Gorman 2024, 11). While Strawson does not say that *all* neurodivergent people are mere moral patients, his focus on the tragic and "hopeless" cases of impaired agency reinforces the stigma around neurodivergent disability. At the same time, the tragedy narrative frames "high-functioning" neurodivergent disabilities as exceptions or "superpowers" possessed by an elite few. Consistent with this trope, some people cite Elon Musk as proof that "some autistic people have extraordinary strengths and talents and can some-

times outperform non-autistic people on certain tasks," especially in the fields of "science and tech" (Shah et al. 2022). This framing creates a dichotomy between the extraordinary, profitable forms of autism on the one hand, and the ordinary, tragic cases on the other. While this dualism may benefit cultural elites like Musk, it oppresses the millions of neurodivergent disabled people whose neurological variations are not culturally valued or celebrated. This is another example of how Strawson's focus on the specter of "disabled misery" perpetuates the dualism between praiseworthy "supercrips" and pitiable "sadcrips."

Gorman adds that while neurotypical/nondisabled people may have good intentions, the unsolicited sympathy that they extend to neurodivergent disabled people is often received as patronizing and paternalistic. Most neurodivergent disabled people do not, in fact, want neurotypical/nondisabled saviours to pity, train, and manage them. Indeed, one of the problems with the kind of "progressive liberalism" that we find in Strawson is that it offers a version of compassionate allyship that merely intensifies oppression. It does so by positioning neurotypical/nondisabled allies as responsible "saviors" who bear the "burden" of healing tragic "cripples," harkening back to the colonial narrative of the "white man's burden" to "civilize the savages." The history of the disability justice movement shows that neurodivergent disabled people are, in fact, capable of liberating ourselves, with or without nondisabled heroes. While political solidarity is appreciated, unsolicited offers of help and healing have a sanctimonious tone, invoking familiar images of messiah figures (St. Pierre 2024).

Virgil Murthy makes a similar point with respect to addiction (2024). Murthy, an addict philosopher, argues that Frankfurtian accounts of responsibility perpetuate patronizing and objectifying stereotypes of addiction by treating addicts as theoretical case studies and objects of intellectual curiosity for non-addicts. Frankfurtians tend to speculate about whether addicts are voluntary delinquents or victims of uncontrollable impulses. Are they mad or bad? Do they deserve punishment or treatment? These armchair speculations, at best, ignore the structural injustices that addicts face on a daily basis, and at worst, lend credibility to oppressive practices like addict overincarceration and sterilization, which are part of a system of eugenic violence. Along the same lines, we can accuse conventional Strawsonians of promoting paternalistic and objectifying stereotypes of neurodivergent disability by using diagnostic categories as fodder for thought experiments and abstract cases studies, thereby overwriting neurodivergent people's testimony. Instead of continuing to riff on Strawson's armchair speculations about "schizophrenics" and "compul-

sives," why not ask us about our agency? Are we hopeless or hopeful? Are we impaired or oppressed? Surely, we know best.

In a forthcoming book on misknowing and flat stories, Jennifer Lackey, who educates and advocates for incarcerated people, explains how second-hand, theoretical accounts of prisoners reduce their inner lives to "flat stories" or stereotypes. Similarly, theoretical representations of neurodivergent people flatten our stories and misrepresent our agency. As Gorman attests, conventional Strawsonians "have flattened the way we perceive a complex landscape for theoretical purposes, but in real life the messy complexity of agential diversity exists nonetheless and our responsibility norms are simply not flexible enough to account for it" (2024: 6). The solution to this ontological flattening is, as both Lackey and Gorman affirm, granting neurodivergent people the right to speak for themselves. In other words, philosophers need to respect the principle of "nothing about us without us," which aims to give neurodivergent disabled people some level of control over how they are publicly represented.

Incorporating these round stories of disability into philosophy is crucial to correcting an epistemic injustice caused by the dominance of abstract theoretical accounts of neurodivergent disability, detached from neurodivergent people's real-world experiences.

II. Other-than-human Animals

In his work on decolonial philosophy, Shyam Ranganathan contends that Western accounts of responsibility, including Strawson's, are part of a colonial tradition of exclusion and epistemic injustice (2024). While Strawson is not solely to blame, it is notable that Strawsonians rarely engage with non-Western perspectives, including Ranganathan's specialisation of South Asian philosophy. This omission perpetuates the myth that moral philosophy is a Western invention, to which South Asian thinkers can contribute nothing of value. This, in turn, negates a rich source of anti-colonial moral philosophy, including critiques of Western imperialism that hold the West responsible for systemic violence toward subjugated groups, including racialised humans, non-human animals, and living ecosystems.

Another upshot of the Western bias in moral philosophy is that it creates an illusion of consensus, making tendentious claims seem intuitive and self-evident. One example of this false consensus is the "commonsense" assumption that responsibility is a property of human beings exclusively. This Strawsonian exemption, however, is not widely held outside of the contemporary

West. Many South Asian philosophers, including Ranganathan, believe that moral responsibility is shared by humans, non-human animals, and interconnected ecosystems like prairies, marshes, and rivers. They argue that the exclusion of non-human species from the responsibility system is a form of human supremacy rooted in colonialism – a system of domination that regards oppressed others, whether human or not, as "uncivilized," lesser, and morally impaired. This exclusionary ethic justifies the use of "impaired agents" as slaves, commodities, and resources. But it is made to appear uncontroversial by an epistemology of domination that silences "external standpoints."

Shelbi Meissner and Andrew Smith affirm this critique, arguing that Western philosophy marginalises Indigenous standpoints that espouse a more inclusive and less anthropocentric definition of responsibility (2024). According to Indigenous environmental feminisms, responsibility is a property of "extended more-than-human kinship relationships," which encompass humans as well as "other-than-human beings, spiritual and abiotic entities, and landscapes" (2024, 14-19). Moral responsibility, as such, is a relational property distributed across networks of relationships that include, but extend far beyond, the human community – a community that is a minority population in the moral domain. Despite this minority status, human beings do by far the most damage to the moral ecology, undermining the ontological basis of responsibility itself, i.e., the kinship relationships of respect and reciprocity that make responsibility, and life itself, possible. Colonialism, then, is one of the greatest existential threats to responsibility, insofar as it constructs asymmetrical relationships through systemic violence, producing vast ecological destruction. How easy is it to be responsible when your house is on fire? This is the "moral community" built by centuries of colonization.

From a settler colonial standpoint, it might be difficult to grasp how other-than-human animals, let alone lakes and streams, could be amenable to the reactive attitudes. After all, we have been taught from an early age to see non-human species as oppositional others. However, other-than-human animals have more in common with us than we tend to think.

Dorna Behdadi offers a revisionary interpretation of Strawson, explaining how canines participate in complex, norm-governed social interactions that involve the exchange of the reactive attitudes. This behaviour makes them inclusive members of the moral community:

> Canid play behavior is an area that is well studied and might prove to be a surprisingly good example of an interaction that involves social norms, as well as expectations, censure, and sanctions that naturally follow interaction that involves

> such norms. Domestic dogs, as well as wolves, coyotes, and other canids, play with one another. During play, canids need to continually assess each other's behavior and intentions and to follow certain play-specific rules (2021, 231).

During play, canids display forms of the reactive attitudes, such as "'trust' or praise" when meeting familiars, regret and remorse after violating a rule, and "surprise" or disapprobation in response to norm violations (2021, 232). Canids also "display appeasement behaviors" that suggest they are "trying to make up, apologize, and/or show remorse," as well as demonstrating "reconciliatory behaviors" that function as "an explanation, acknowledgement, or excuse" (2021: 233) Even if we accept a capacity-based interpretation of Strawson, we should grant, in light of behavioral science, that many other-than-human animals share our capacity to participate in the moral community and respond to the reactive attitudes.

This recognition of nonhuman personhood, which runs afoul of the West's instrumental treatment of animals – who are commodified and exploited through factory farming, animal experimentation, and other exploitative industries – is normative in many Indigenous communities, including the Xeni Gwet'in First Nations people of British Columbia, who have lived with domesticated horses since before European colonization. The Xeni Gwet'in people describe horses, both domesticated and wild, as "family members," "neighbours," and members of the community (Bhattacharya & Slocombe 2017: 8). Horses are a central part of the tribe's cultural identity and customs. Their kinship with the local horses grounds both positive and negative reactive attitudes. The Xeni Gwet'in people value, praise, and take pride in local horses for their "toughness, endurance, and speed" (2017: 7). They admire horses who are "sure-footed" and "wildlife savvy," "know the local terrain," and "think for themselves" (ibid). These qualities elicit praise, approbation, trust, and other forms of positive regard. By the same token, horses who lack these qualities may be seen in a negative light. If they are unreliable, difficult, or short-tempered, the horse may elicit disapproval, disappointment, or hurt feelings from a steward who has raised the horse well, in accordance with their stewardship responsibilities. Kinship relationships ground a reciprocal expectation of cooperation, generosity, and respect, and whether this expectation is fulfilled or disappointed decides what reactive attitudes are appropriate between horse and human.

Notably, the basis for these attitudes in kincentric tribes is not capacity (e.g., language, moral reasoning), but rather kinship, which is a relational property distributed across the web of life. Hence, not merely horses and

wolves, but also rivers and streams are fellow persons, amenable to the reactive attitudes. Kyle Whyte explains how many Indigenous communities have an "emotional-laden relationship" to "features of the land (like rivers or mountains)" along with "natural interdependent collectives," and this relationship is grounded in reciprocal responsibilities (2014: 602). For example, "a community may have a responsibility to care for salmon habitat; salmon, in turn, may provide food and support for other species" (Whye 2014: 603). If the community fulfils its responsibilities toward the salmon, they are entitled to expect support in return. They may be grateful if this expectation is surpassed, or disappointed if it is frustrated. In this way, the reactive attitudes play a role in structuring and regulating relationships of reciprocal responsibilities between humans and nonhuman persons. Humans, fish, rivers, and other interdependent collectives form a moral network shaped by reciprocal responsibilities, which ground positive and negative reactive attitudes.

White explains further how Anishnaabe people have moral expectations of lakes and streams:

> Many Anishinaabe people value water greatly, which arises from the Anishinaabe creation story in which water is considered to play the role of a source and supporter of life. In this role, water mediates interactions among many living beings on the earth. Consequently, water is a considered a relative that has responsibilities to give and support life… Humans, in turn, have responsibilities to care for and respect water; they must especially do things that encourage water's life-giving force. Ceremonies are structured to remind people of their connections to water, and bodies of water are considered to have their own unique personalities (2014, 605).

If humans fulfil their responsibilities to the water, they are entitled to expect reciprocal benefits. Depending on whether this expectation is fulfilled, positive or negative regard is in order.

Many Indigenous communities also have emotionally-laden, norm-governed relationships with plants. This includes Wabanaki cultures, in which "the responsibilities surrounding berry plants have intrinsic value because they are integral to customs and rituals and establish part of the cultural status of Wabanaki women… Thus, an entire system of responsibilities is embedded in and permeates everything about the berry plants" (Whyte 2014, 603). Wabanki people expect berry plants to fulfil their expectations if they have fulfilled their own duties of care, stewardship, and respect toward the plants. Whether and to what extent these duties are fulfilled decides what types of reactive emotions are fitting.

Whyte emphasizes that colonialism has disrupted these reciprocal relationships, making it increasingly difficult for persons to fulfil their responsibilities to each other. For the Wabanki people, "climate change may affect the range, quality, and quantity of species like berries, making it more difficult or even impossible for tribal members to assume the responsibilities they perceive themselves to have toward those species" (ibid). As colonial systems continue to encroach on every ecosystem on the planet, especially those cultivated and cared for by oppressed groups, it becomes harder for *anyone* to fulfil their responsibilities to others. This reinforces that climate injustice remains amongst the greatest threats to responsibility – a threat that is neglected in settler theories of responsibility that focus on individual human capacities. Individuals exist within networks of interdependence and vulnerability that ground reactive emotions, justifying moral attitudes not only between humans but across the web of life.

III. Young Children

The philosophical literature on children's responsibility is, to my knowledge, relatively scarce. However, this does not mean that critiques of adult supremacy are any less valuable than critiques of human and neurotypical supremacy. On the contrary, these critiques are a crucial part of the concerted effort against exclusionary moral philosophies, which tend to view neurotypical human capacities as the sole basis for moral responsibility. Fortunately, critiques of colonialism and ableism contain tacit critiques of adult supremacy.

Indigenous environmental feminisms, for example, seem to imply, even if they do not explicitly state, that young children are moral agents. This follows from the claim that extended human and more-than-human kinship networks confer moral agency, and these networks include young children. Thus, young children are moral persons with rights and responsibilities. It is reasonable to express gratitude, approval, disapprobation, disappointment, and other moral emotions to young children, and to be amenable to the same attitudes in return.

The neurodiversity model supports the same conclusion for similar reasons. It rejects the notion that adult neurotypical capacities are required for responsibility, and, by extension, that neurotypical adults are the sole proprietors or exemplars of responsible agency. If neurotypicality is not required for responsibility, then children can be responsible agents, too.

Like Indigenous feminists, critical disability theorists tend to reject the capacities view in favor of a relational model. Eva Fedder Kittay, for example, views interdependence, vulnerability, and emotional ties as the basis of moral personhood. Since we all depend on interpersonal relationships, we are all persons, with inviolable rights and responsibilities. Kittay rejects neoliberal (individualistic) models that define personhood as a property of autonomous, self-sufficient individuals, describing these models as "morally repugnant," and comparable to "earlier exclusions based on sex, race, and physical ability" (2005: 100). Licia Carlson similarly argues that treating intellectual disability as a limiting or "marginal case" of personhood is an offensive form of "cognitive ableism" (2021, 74), and this argument can be extended to neurodivergent disability. The relational definition of personhood is meant to incorporate cognitively disabled people into the definition of personhood, but it can also be used to validate the personhood of children, who are a vulnerable and dependent social group.

Stephanie Jenkins notes that the philosophical literature on responsibility tends to regard "animals and people with disabilities" as mere "moral patients" or "non-contributing dependents" (2024, 378). This exclusionary definition of personhood – which resembles Strawson's understanding – is useful to authoritarians: "For example…, Nazi ideology encouraged a constriction of the moral community. A derealization of the other impeded the affective moral responses of German citizens" (Jenkins 2024, 379). The constriction of the moral community, in other words, is not just morally significant but also politically expedient: it helps cultural elites control and exploit oppressed populations – those excluded from the dominant culture. As a bulwark against authoritarianism, Jenkins proposes a "precautionary principle of moral status," which "begins with the assumption that animate life, whether human or nonhuman, abled or disabled, deserves moral concern" (2024, 384). This principle should help us avoid sliding further into the authoritarian political regime from which canonical Western philosophy emerged. [3] Expanding the moral community beyond what Strawson envisioned could, in other words, contribute to epistemic and political justice.

The precautionary principle is meant to incorporate neurodivergent/disabled and nonhuman beings into the scope of personhood, but it also encompasses young children, who are equally subject to oppression, though

[3] I am thinking here of canonic moral philosophers like Kant, who defended white supremacy, Locke, who defended the colonization of "savage, primitive societies," and Heidegger, a card-carrying Nazi.

this is less widely recognized. Yet legal minors are highly susceptible to violence and abuse due to their situation of political disenfranchisement. According to the U.S. Department of Justice, American children experience and witness more violence than adults (2020). Nonetheless, the U.S. refuses to sign the United Convention on the Rights of the Child, which would protect children from "physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, at school or by a parent or legal guardian" (Congressional Research Service 2015). The U.S. reserves the right of adults to batter children, marry children, rape children in the context of marriage, and perpetuate other types of abuse that would be illegal if done to an adult. If children are mere moral patients, they can do nothing to address their circumstances of ageist oppression. Instead, they must wait for adults to act on their behalf – a highly unlikely scenario. If adults are the main perpetrators of violence against children, how can they be expected to protect them? As Frederick Douglass said, "power concedes nothing [willingly]" (1857).

Defining children as moral patients, devoid of moral agency or autonomy, is at odds with the mission of the Youths Rights Movement, which holds that young people are denied basic, inviolable rights due to ageist discrimination; "These rights include the right to be full participants in our representative democracy through voting, the right to privacy, the right to be free from physical punishment, the right to make decisions about our own lives, the right to be outdoors, the right to prove ourselves, and the right to receive the same amount of respect as anyone else" (National Youth Rights Association 2024). Ageism against youths "is a form of discrimination," but "unlike discrimination based on race, gender, religion, sexual orientation, ability, or age (at least for older people), ageism against the young (sometimes called adultism) is both legal and common" (ibid). Denying that children are moral agents with rights and responsibilities conflicts with the Youth Rights Movement's mission to empower and enfranchise young people.

It is also at odds with the historical record of youth activism, which shows that youths have always exercised political agency by participating in causes that are often thought of as "mature" and "adult." In 1963, almost 1,000 children were jailed after skipping school to participate in the Birmingham Freedom Movement to desegregate public schools" (Vole 2024). Young people attended meetings of the NAACP Youth Council, participated in sit-ins, and attended rallies throughout the 20th Century (Library of Congress). Today, many young people skip class to protest climate change (Fridays for Future), school shootings (The National School Walkout), and other adult-caused injustices that disproportionally affect children. The U.S. Institute of Peace affirms that

"recent history is replete with examples [of youth activism] — mass movements in Iran, Hong Kong, Sudan, Lebanon, Algeria and others have all drawn strength from major swells of determined youth mobilization" (Cebul 2023). Yet while "young people help nonviolent campaigns succeed…, youth, [especially girls], do not share equally in the spoils of victory" (ibid). Young people constantly exercise their moral agency, but rarely receive the moral respect and recognition they deserve.

This does not mean that children do not deserve special protections. Like any oppressed group, youths are entitled to constitutional, legal, and social protections based on their situation of oppression. This is compatible with recognizing their rights and responsibilities. [4]

IV. Ameliorative Responsibility

In this paper, I outlined and expanded on some recent interdisciplinary critiques of Strawson's "commonsense" exclusion of non-neurotypical people, other-than-human animals, and young children from the moral community. These critiques, if nothing else, challenge the intuitiveness and self-evidence of the capacities criterion, which is grounded in the pretheoretical intuitions of a specific standpoint – for Strawson, that of a privileged, white, cisgender, male professor at Oxford university in the mid-to-late 20th century. To a modern audience, Strawson's confident pronouncements about how we should treat "the insane" are likely to rankle. Marginalised philosophers are casting doubt on these intuitions, leveraging their own standpoints to defend a more inclusive and intersectional model of responsibility. These critiques help us to not only understand the world of responsibility, but also, and more importantly, to change it.

The goal of changing the world is facilitated by non-ideal theory, which focuses our attention on structural injustices and the testimony of those most affected by them. Rather than imagining theoretical ideals of responsibility from our philosophical armchairs, non-ideal theory instructs us to identify and respond to real-world injustices. This raises practical questions: How should we respond? What should the future look like? These questions are best answered by those with expertise in navigating and surviving oppression: oppressed

[4] This qualification is important because I do not want readers to think that youth suffrage entails repealing laws protecting children from child labor, sexual exploitation, and other adultist crimes that should remain illegal.

people, whose narratives and cultural traditions can provide guidance on how to address crises like climate injustice, overincarceration, forced sterilization, and segregation.

The attempt to use responsibility as a tool of social justice aligns with what Sally Haslanger refers to as an "ameliorative" methodology, which seeks to identify structural injustices and resolve them. Ameliorative conceptual engineering goes hand-in-hand with non-ideal theory's focus on structural injustices that can be addressed. Haslanger believes that when we attempt to define a concept, whether it be race, gender, autonomy, or responsibility, we should not rely solely on intuitions or conventions, but should "develop [a] concept that would help us achieve [our] ends." (2005, 11). If our aim is to mitigate ableism, address colonial violence, liberate nonhuman animals, and address other intersections of oppression, then we should select a concept of responsibility tailored to these purposes – a concept that is not merely revisionary but revolutionary; not merely social but political; not merely descriptive but disruptive. Neurodiversity theory, Indigenous feminisms, and youth liberation activism support an ameliorative approach to responsibility that transforms the system from the ground up.

It is notable that many (but not all) of the arguments presented in support of an ameliorative approach view responsibility as a function of relationships rather than capacities. While much more could be said about the relational model – and has been said elsewhere (e.g., Young 2011, Walker 1998, Held 2006, Ziegler 2016) – my present aim is not to vindicate the relational view, but merely to introduce original critiques of Strawson from a variety of cultural and disciplinary standpoints, thereby challenging the illusion of consensus around the philosophical understanding of responsibility. My goal, in other words, is to problematize – or, to use Judith Butler's term, "trouble" (2002) – the philosophical notion of responsibility, encouraging broader intersectional debate about its social role. New critical perspectives on Strawson enhance epistemic justice and diversity, offering a more nuanced understanding of what responsibility is and could be. [5]

When using an ameliorative lens, we should consider diverse historical, cultural, and disciplinary perspectives, keeping in mind what we want responsibility to do for us, both now and in the future. What kind of world do we

[5] In *What Love Is: And What It Could Be* (2017), Carrie Jenkins proposes an ameliorative theory of love, which critically examines its past and present functions, and asks us to reflect on what we want love to do for us now and in the future, in light of its history. This is the method that I am proposing for moral responsibility.

want to live in: one that sees disabled people as "hopeless," helpless, "risks to be avoided," or one that recognizes disabled people as full members of the moral community? A world that sees other species as objects and commodities, or one that sees them as kin and neighbours? A world that gives adults full control over their children's lives, or one that recognizes young people's rights and responsibilities? Anti-oppressive discourses provide a counterpoint to the conventional wisdom, guiding us toward a more inclusive, intercultural, and reciprocal future.

Works Cited

Bhattacharyya, J., and S. Slocombe. 2017. "Animal Agency: Wildlife Management from a Kincentric Perspective." *Ecosphere* 8(10): 8.

Behdadi, D. 2021. "A Practice-Focused Case for Animal Moral Agency." *Journal of Applied Philosophy* 38(2): 231.

Bureau of Labor Statistics. 2020. "Persons with a Disability: Labor Force Characteristics—2019". https://www.bls.gov.

Butler, J. 2002. *Gender Trouble*. Routledge.

Carlson, L. 2021. "Why Does Intellectual Disability Matter to Philosophy?: Toward a Transformative Pedagogy." *Philosophical Inquiry in Education* 28(2): 72-82.

Cebul, M. D. 2023. "Balancing Risk and Reward." United States Institute of Peace. https://www.usip.org.

Cherry, M., and E. Schwitzgebel. 2016. "Like the Oscars, #PhilosophySoWhite." *LA Times*, March 4. https://www.latimes.com. Retrieved December 6, 2024.

Congressional Research Service. 2015. "The United Nations Convention on the Rights of the Child". https://crsreports.congress.gov.

Dembroff, R. 2020. "Cisgender Commonsense and Philosophy's Transgender Problem." *Transgender Studies Quarterly* 7(3): 403.

Farley, M., K. J. Cottle, D. Bilder, J. Viskochil, H. Coon, and W. McMahon. 2018. "Mid-Life Social Outcomes for a Population-Based Sample of Adults with ASD." *Autism Research* 11(1): 142-152.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Frye, M. 1983. *Politics of Reality: Essays in Feminist Theory*. Crossing Press.

Gorman, A. 2024. "Against Neuronormativity in Moral Responsibility." *Feminist Philosophy Quarterly* 10(1/2).

Haslanger, S. 2005. "What Are We Talking About? The Semantics and Politics of Social Kinds." *Hypatia* 20(4): 10-26.

Held, V. 2005. *The Ethics of Care: Personal, Political, and Global*. Oxford University Press.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33(2-3): 61-83.
Holroyd, J. 2024. "The Distortions of Oppressive Praise." *Feminist Philosophy Quarterly* 10(1/2): 1-30.
Jenkins, C. 2017. *What Love Is: And What It Could Be*. Hachette UK.
Kittay, E. F. 2005. "At the Margins of Moral Personhood." *Ethics* 116(1): 100-131.
Lackey, J. 2024. "Misknowing and Flat Stories." Presented at the North American Society for Social Philosophy Conference, July 12, 2024.
Library of Congress. "Youth in the Civil Rights Movement". https://www.loc.gov.
McGeer, V. 2019. "Scaffolding Agency: A Proleptic Account of the Reactive Attitudes." *European Journal of Philosophy* 27(2): 301-323.
McKenna, M. 2012. *Conversation & Responsibility*. OUP USA.
Meissner, S., and A. Smith. 2024. "Climate Crisis as a Relational Crisis." *Feminist Philosophy Quarterly* 10(1/2): 1-30.
Mills, C. W. 2017. *Black Rights/White Wrongs: The Critique of Racial Liberalism*. Oxford University Press.
National Youth Rights Association. "What Are Youth Rights?" https://www.youthrights.org.
Oliver, M. 2013. "The Social Model of Disability: Thirty Years On." *Disability & Society* 28(7): 1024-1026.
Pateman, C. 1988. *The Sexual Contract*. Stanford University Press.
Prison Policy Initiative. 2022. "Disability". https://www.prisonpolicy.org.
Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Ranganathan, S. 2024. "Reason and Solidarity with Persons Against White Supremacy and Irresponsibility: A South Asian Analysis." *Feminist Philosophy Quarterly* 10(1/2): 1-31.
Shah, P., L. Hargitai, and L. A. Livingston. 2022. "Elon Musk: How Being Autistic May Make Him Think Differently." *The Conversation*. https://theconversation.com.
Shoemaker, D. 2017. "Response-Dependent Responsibility; or, A Funny Thing Happened on the Way to Blame." *Philosophical Review* 126(4): 481-527.
Simplican, S. C. 2016. *The Capacity Contract: Intellectual Disability and the Question of Citizenship*. University of Minnesota Press.
Strawson, P. F. 1963. "Freedom and Resentment." Open-access source, unpaginated.
Talbert, M. 2020. "Moral Responsibility." In *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), edited by Edward N. Zalta. https://plato.stanford.edu.
Tremain, S. 2024. "When Moral Responsibility Theory Met My Theory of Disability." *Feminist Philosophy Quarterly* 10(1/2): 1-30.
Tremain, S. 2021. "Dea, Data, and the Disabling Canadian University." *BIOPOLITICAL PHILOSOPHY* (blog), November 8. https://biopoliticalphilosophy.com. Retrieved December 2, 2024.
Tremain, S. 2017. *Foucault and Feminist Philosophy of Disability*. University of Michigan Press.

U.S. Department of Justice: Office of Justice Programs. 2020. “Children Exposed to Violence”. https://www.ojp.gov.
Vargas, M. 2004. “Responsibility and the Aims of Theory: Strawson and Revisionism.” *Pacific Philosophical Quarterly* 85(2): 218-241.
Vole, A. 2024. “Birmingham Children’s Crusade.” *Encyclopaedia Britannica*. https://www.britannica.com.
Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. Harvard University Press.
Walker, M. U. 1998. “Ineluctable Feelings and Moral Recognition.” *Midwest Studies in Philosophy* 22: 62-81.
Whyte, K. P. 2014. “Indigenous Women, Climate Change Impacts, and Collective Action.” *Hypatia* 29(3): 599-616.
Young, I. M. 2011. *Responsibility for Justice*. Oxford University Press.
Ziegler, Z. 2016. “A Relational Theory of Moral Responsibility.” *Prolegomena: Časopis za filozofiju* 15(1): 71-88.