Undergraduate Philosophy Journal *of* Australasia

Volume 8
August 2026

UPJA

Founded in 2019, the *Undergraduate Philosophy Journal of Australasia* (UPJA) is the first undergraduate philosophy journal run by students from Australasia. We publish one volume and host two conferences annually and interview philosophers with a substantial connection to Australasia. We aim to be an inclusive and diverse journal and welcome submissions from undergraduates (and recent graduates) worldwide, on any philosophical topic, so long as the author attempts to make a substantive contribution to contemporary philosophy. Submissions from women and other members of underrepresented groups in philosophy, including those for whom English is not their first language, are particularly encouraged.

Undergraduate Philosophy Journal of Australasia – UPJA

@UPJA_Philosophy

upja.philosophy@gmail.com

www.upja.online

ISSN: 2643-3146

Editors

Ava Broinowski, University of Sydney

Ava completed her Honours year at the University of Sydney in June 2026 (First Class & the University Medal). Her thesis aimed to uncover a notion of the unconscious in the writings of David Hume, particularly from within the account of habit in the *Treatise of Human Nature*, which guides patterns of thinking that precede reflective deliberation. Her interests include the history of philosophy, embodied cognition, social epistemology, hermeneutics, aesthetics, genealogy and ecology.

Winny Li, University of Wollongong

Winny completed her undergraduate degree at the University of Wollongong focusing on the intersection between theoretical thought and real-world application. Winny's philosophical work centres on epistemology, political philosophy, and identity theory, with a strong interest in interdisciplinary approaches that connect diverse fields of inquiry.

Charlotte Carnes, Australian National University

Charlotte graduated with First Class Honours in philosophy from the Australian National University in November 2025. Her Honours thesis explored how the dominant Western imaginary of pregnancy, in which the relationship between a woman and foetus is equated to that between two individuals, informs an understanding of abortion as a strictly moral question. More broadly, her philosophical interests lie in feminist philosophy, bioethics, moral and political philosophy, and metaethics.

Jacob Ritz, University of Queensland

Jacob is a PhD student in the School of Languages and Cultures and casual academic in mathematics and German studies at the University of Queensland. He is also the associate editor of the undergraduate journal Exordium. His research interests concern the history of nineteenth- to twenty-first-century French and German philosophy. His PhD thesis develops an intellectual history of Alain Badiou’s philosophy, examining how it is informed by nineteenth- and twentieth-century German metaphysics.

Referees

Georgia Farmer, University of Queensland

Harley Harris, University of Melbourne

Conor Jedam, University of Queensland

Paris Kastanias, University of New South Wales

Matthew DeKock, University of Toronto

Ruiwen Guo, Australian National University

Moeka Koyama, Harvard University

Haemish Lander McBride, University of Queensland

Victoria Lawson, University of Queensland

Andrew McGuiness, University of Sydney

Sean Ethan Montoya-Capisanan, Pontifical and Royal University of Santo Tomas

Nathaneal Navarro, University of the Phillipines

Hyerim Park, University of Sydney

Robert Powell, Macquarie University

Enora Richard, Australian National University

Phoebe Sampson, University of Queensland

Forest Scaife-Elliott, University of Queensland

Tessa Popple, University of Queensland

Jordyn Yeung, University of Queensland

Editors' Notes

The 2026 winter volume of the UPJA presents five papers with far-reaching scope in contemporary philosophy. Over the past year, the UPJA saw the return of the annual conference *in-person*, which took place at the University of Sydney. We are both delighted to witness the UPJA community grow, and to have contributed to the important tradition of sharing ideas at the undergraduate level.

This volume contains five stellar papers, which happen to pertain particularly to analytic concerns in epistemology, philosophy of mind, and ethics. Piet Fritz Pankratz contributes a thoroughgoing account of the relation between corporate agents and their members building on the work of K. Hess. He contributes to a notion of the complementary, rather than conflicting, nature of *grounding* and *emergence* relations. Enora Richard assesses the limited applicability of models of responsibility that would treat beneficiaries of climate-polluting activities as obligated to assist in instances of climate-induced-hardship. Tessa Popple discusses the functionalist presuppositions behind A. Clark and D. Chalmers' Extended Mind Hypothesis, and offers a critique of the theory precisely on functionalist grounds. Bowen Northrop considers the relation between aesthetic self-expression and M. Fricker's hermeneutical injustice, centering obstructions of non-binary self-expression as an instance of a more broadly applicable kind of hermeneutical injustice. Finally, Matthew DeKock, informed by C. Misak, appraises F.P. Ramsey's theory of belief, questioning how it is that a dispositionalist theory which identifies actions *with* beliefs can accommodate the familiar phenomenon wherein one acts *against* their better beliefs (akrasia). Overall, though we run submissions with a generic theme, Volume 8 has come to feel especially relevant to and reflective of its times. We hope our readers will appreciate the insights of these excellent undergraduate papers.

Special thanks are owed to the previous Editors-in-Chief, Paul Kim and William Smith, for their consistent mentorship and contribution to the UPJA project. This semester, we welcomed Charlotte Carnes and Jacob Ritz as associate editors. They have imparted considerable vitality and rigour towards realising this volume, and we are so excited to pass the baton. Finally, we are immensely thankful to our team of referees, who, due to a considerable volume of submissions, contributed plenty of effort and discernment towards refining every submitted paper.

Ava Broinowski and Winny Li

Editors-in-Chief

August 2026

Table of Contents

* Winner of Best Paper

† Winner of Best Paper (Member of an Underrepresented Group in Philosophy)

Undergraduate Philosophy Journal of Australasia www.upja.online
Vol. 8, August 2026

Grounded Corporate Agents with Emergent+ Capacities

Piet Fritz Pankratz[1]

University of Vienna

Abstract

This paper clarifies the metaphysical relation between corporate agents and their members under Kendy Hess's holist account of corporate agents. While Hess emphasizes that corporate agents depend for their existence on the structured participation of their members, the precise nature of this dependance relation remains insufficiently articulated. I argue that corporate agents are *non-reductively grounded* in their members: they exist in virtue of the structured participation of individuals within enduring social organizations. Grounding thus explains the ontological dependence of corporate agents on their members. However, on Hess's view, corporate agents possess capacities and properties that can only be attributed to the organized whole. The grounding relation leaves these distinctive capacities and properties insufficiently explained. To address this, I introduce a modified notion of *emergence* ("emergence+"), according to which corporate agents' capacities and properties depend upon, yet are not reducible to, their members. I further argue that grounding and emergence+ are complementary rather than competing relations. Taken together, grounding and emergence+ offer a more complete account of the corporate agent–member relation and clarify central but underdeveloped metaphysical commitments of Hess's holism.

[1] *Piet Fritz Pankratz is an undergraduate philosophy student at the University of Vienna. His research interests include philosophy of science, social ontology, and Logical Empiricism (especially Rudolf Carnap, who would likely have disapproved of this paper).*

§ 0. Introduction

Moral responsibility is frequently ascribed not only to individuals but also to collectives. In particular, it is often attributed to a subclass of collectives: corporations. When, for example, a product malfunctions and causes harm, the locus of blame typically falls on the corporation itself (e.g., Apple) rather than on any individual employee. Here, we treat the corporation as the bearer of responsibility. Doing so is typically not merely for epistemic reasons. That is, it is not simply that people blame the corporation because they lack knowledge of which specific employee(s) contributed to the malfunction. Rather, there appears to be a genuine sense in which blame is directed at the corporation as such. This practice seems to presuppose the existence of corporate agents, for if we are to make sense of corporate responsibility, we must posit agents to whom such responsibility can be attributed. By "corporate agents," I refer, broadly, to businesses such as Apple or Ströck-Brot GmbH (a well-known Austrian bakery chain). While such agents may vary in size and complexity, my focus is specifically on corporate agents that exhibit structured decision-making capacities and are therefore more plausible candidates for moral responsibility.

When we posit corporate agents to attribute responsibility to them, we are immediately faced by a tension: while a corporate agent is composed of individual agents (i.e., its members), its status as a genuinely responsible entity seems to require a degree of ontological independence from those individuals. Intuitively, the relation between a corporate agent and its members is deeply entangled; yet any robust account of corporate agents must also secure a measure of autonomy for the corporate agent. It is somewhat surprising that the corporate agent–member relation has not yet been examined with the degree of explicitness that its significance would appear to warrant. As a result, the nature of this relation remains comparatively under-theorized and lacks conceptual clarity. The aim of this paper is to elucidate the corporate agent–member relation. I will do so by drawing on two concepts from analytic metaphysics: grounding and emergence.

More broadly, I seek to address the ontological challenges that confront the notion of a corporate agent, in line with the work of Kendy Hess.[2,3] These challenges are especially pressing because, as Hess rightly notes, the corporate agent's metaphysical underpinnings remain significantly underdeveloped, despite being one of the most common points of critique against collectivist approaches.[4] Many critics

---

[2] While questions concerning corporate agents frequently give rise to epistemic challenges, for example, regarding how we might identify, predict, or model their behavior, such issues lie beyond the scope of this paper.

[3] Hess, Kendy (2018a) 'Does the Machine Need a Ghost? Corporate Agents as Nonconscious Kantian Moral Agents', *Journal of the American Philosophical Association* **4**; Hess, Kendy (2018b) 'The Peculiar Unity of Corporate Agents', in Kendy Hess, Victoria Igneski, and Tracy Isaacs, eds., *Collectivity: Ontology, Ethics, and Social Justice*, Rowman and Littlefield; Hess, Kendy (2020) 'Assembling the Elephant: Attending to the Metaphysics of Corporate Agents', in Deborah Tollefsen and Saba Bazargan-Forward, eds., *The Routledge Handbook of Collective Responsibility*, Routledge; Hess, Kendy (2025) 'Re-bunking Corporate Agency', *Inquiry: An Interdisciplinary Journal of Philosophy* **68**: 979-99.

[4] Hess, 'Re-bunking Corporate Agency', 980.

of the collectivist position have objected to it on the grounds that it entails what they perceive as ontologically extravagant or metaphysically peculiar commitments. As Hess observes, these critics frequently invoke an "aura of weirdness" surrounding collectivist claims.[5] In particular, corporate agents are often portrayed as "ghostly," "incorporeal," "mysterious," "magical," "spooky," "hovering" entities.[6]

This paper shares Hess's aim of developing a "metaphysically robust" account of corporate agents, one capable of withstanding the common charge that collectivist theories lack an ontological foundation.[7] Following Hess, I do not discard the influential accounts of corporate agency advanced by Peter French and Christian List and Philip Pettit (hereafter collectively referred to as "FLiP").[8] Rather, the goal is to enhance their plausibility by addressing and moving beyond certain restrictive assumptions embedded within their frameworks.[9] There is a general tendency within FLiP's accounts of corporate agency to privilege corporate agents' officially recognized practices, those codified or institutionally sanctioned, while neglecting informal, unofficial, or unwritten practices that nonetheless play a significant role in shaping the corporate agent. I submit that the project initiated by Hess may be understood as an effort to de-idealize FLiP's accounts, thereby rendering them more applicable to a broader range of corporate phenomena.

It should be emphasized that the present paper does not undertake a systematic defense of Hess's holist account of corporate agents. Rather, its aim is interpretive and constructive: to develop conceptual resources that clarify and further articulate the metaphysical commitments implicit in Hess's framework. Although such clarification may bear on the ultimate plausibility of the view, a systematic defense of holism would require a separate argument and lies beyond the scope of this paper. Accordingly, Hess's conception of the corporate agent is taken as a working assumption throughout.

Ultimately, I argue that the relation between corporate agents and their members is best understood as having two irreducible metaphysical dimensions: corporate agents are grounded in their members insofar as they depend on them for their existence, yet corporate agents' capacities are emergent from their members insofar as those capacities can only be attributed to the organized whole.

I proceed as follows: Section 1 articulates the theoretical framework within which the argument is developed, outlines Hess's holism, and introduces grounding and emergence; Section 2 argues that corporate agents are grounded in their members; Section 3 argues that corporate agents' capacities and properties are emergent from

[5] Hess, 'Re-bunking Corporate Agency', 994.
[6] For a catalogue of these critiques: Hess, 'Re-bunking Corporate Agency', 980 fn. 2.
[7] Hess, 'Does the Machine Need a Ghost?', 68; Hess, 'The Peculiar Unity of Corporate Agents', 36; Hess, 'Assembling the Elephant', 119; Hess, 'Re-bunking Corporate Agency', 984.
[8] The most important works on corporate agency by FLiP include: French, Peter (1984) *Collective and Corporate Responsibility*, Columbia University Press; List, Christian and Philip Pettit (2011) *Group Agency: The Possibility, Design, and Status of Corporate Agents*, Oxford University Press.
[9] Hess, 'Re-bunking Corporate Agency', 985–87; Hess, 'Assembling the Elephant', 122.

their members; Section 4 reconciles the grounding and emergence relations developed in the preceding sections; Section 5 concludes.

§ 1. Theoretical Framework

This section establishes the theoretical framework of the argument, by outlining Hess's holist account of corporate agents and introducing the concepts of grounding and emergence independently of Hess's work.

1.1 Hess's Holist Account of Corporate Agents

The holist is committed, on Hess's account, to two central theses: first, that properly organized collectives of persons can instantiate a metaphysically distinct entity; second, that such entities possess irreducible properties and capacities, notably including "the capacity for a highly sophisticated form of rational agency".[10] On this account, a corporate agent is not to be reductively identified with the intentional states, decision procedures, or the agency of its individual members alone. Rather, it is unified by the underlying *social structure* that merges its members into a coherent whole, which enables it to exhibit certain complex capacities (e.g., the capacity to deliberate and act). In what follows, I adopt Hess's demarcation of membership, according to which all individuals embedded in the workforce of the social structure qualify as members. This demarcation is intended to delineate membership boundaries in a principled way: "support staff who are embedded in the social structure are members, e.g., while outside consultants are not, however essential and effective their influence may be."[11] For Hess, this demarcation follows from her claim that participation in the relevant social structures is immersive rather than merely episodic.[12] It is this immersive mode of participation that underwrites the robust ontological status she attributes to corporate agents and, in turn, allows them to be understood as material entities. On Hess's view, this materiality provides "an outer limit to the subvenient base," leading to the above demarcation.[13]

A further pivotal element of Hess's account is what she terms "[r]e-membering the corporate agent".[14] Hess aims to foreground the role of corporate members within the corporate agent, countering interpretations, such as those she attributes to FLiP, that risk "dis-membering" the corporate agent and thus undermine its plausibility.[15] In particular, FLiP detach the corporate agent from its members in a way that makes it implausible to think that a corporate agent could ever materialize from its members. Contrastingly, Hess's re-membering suggests a metaphysically

[10] Hess, 'Re-bunking Corporate Agency', 979.
[11] Hess, 'Assembling the Elephant', 122.
[12] Hess, 'Assembling the Elephant', 122.
[13] Hess, 'Assembling the Elephant', 122.
[14] Hess, 'Re-bunking Corporate Agency', 984.
[15] Hess, 'Re-bunking Corporate Agency', 982–84; 987–88.

significant relation between the corporate agent and its members: the members constitute an integral part of the corporate agent.[16] The present inquiry aims to specify this relation.

Turning to the social structure mentioned above, Hess emphasizes that it is this structure which unifies the disparate members into a single ontological whole.[17] Such unity is possible because the structure is understood as "a network of social relationships internal to the corporate agent which constrain the actions of the members."[18] Through these relationships, the structure organizes the members' thought, communication, and actions. In more concrete terms, the members are provided with "templates that favour certain forms of action, coordination, justification, and axiology among other things."[19] While this type of social structure is necessary for the existence of the corporate agent, Hess denies that it is sufficient: the "active participation of the members" is also required for its formation.[20] Accordingly, she insists that the corporate agent is not identical to the social structure itself. Hess further characterizes this structure as broad and deeply embedded, affording a more comprehensive and resilient integration of individuals than alternative accounts, such as those offered by FLiP, can provide.[21] Participation within this structure entails more than discrete actions such as voting or role fulfillment; it requires an "immersive" kind of engagement.[22] Here, "immersive" denotes a mode of participation in which members are substantively and holistically integrated into the social structure, such that their agency is embedded within and constituted by the ongoing practices and relations that sustain the corporate agent.

Building on these considerations, Hess advances a metaphysically substantive view of corporate agents as *material entities*.[23] She asserts that, analogous to individual human agents, the intentionality of corporate agents is rooted in their "material aspect".[24] However, unlike individual agents, the material aspect of corporate agents is distributed across members and unified through consent.[25] This consent-based unity should not be understood as a merely contractual relation. Rather, Hess argues that the distributed material aspect of a corporate agent is held together by members' ongoing willingness to occupy and enact organizational roles within the corporate structure. The corporate agent is therefore unified not by physical contiguity, as in the case of an individual organism, but by the sustained participation of individuals who submit themselves to the constraints and

[16] Interestingly, Hess appears to take an initial step toward capturing this relation in her 2020: "it reveals corporate intentionality and agency as [...] more deeply *rooted in* the full membership and more *independent of* any individual member than the current accounts allow." Hess, 'Assembling the Elephant', 124, my emphasis.
[17] Hess, 'Re-bunking Corporate Agency', 987.
[18] Hess, 'Re-bunking Corporate Agency', 985.
[19] Hess, 'Re-bunking Corporate Agency', 985.
[20] Hess, 'Re-bunking Corporate Agency', 989.
[21] Hess, 'Assembling the Elephant', 122.
[22] Hess, 'Assembling the Elephant', 122.
[23] Hess, 'Re-bunking Corporate Agency', 988–994.
[24] Hess, 'Re-bunking Corporate Agency', 988–990.
[25] Hess, 'Re-bunking Corporate Agency', 988.

expectations embodied in the structure. Consent functions as the mechanism through which a dispersed collection of persons is formed into a single organized entity.

Hess defends this claim by drawing an analogy with other scattered yet uncontroversially material entities.[26] The fact that a corporate agent's members are spatially distributed and connected through voluntary participation does not render the entity immaterial or mysterious. What matters is that the members are integrated by a stable structure that coordinates their activities and preserves the unity of the whole over time. The corporate agent therefore possesses a material aspect rooted in its members, even though the bonds unifying those members are social rather than physical. Importantly, on Hess's view, dependence on consent bears only on *how* the corporate agent comes into being, not on whether there is a corporate agent in the first place.

More broadly, Hess's account reveals corporate agents to be "richer, more complex, and far more autonomous vis-à-vis corporate members than the other theories can account for."[27] In this way, Hess's holism enables us to accommodate the overwhelming complexity of corporate agents into a broadly unified but still multi-component framework. Consider Apple. It comprises thousands of employees across departments, technologies, and geographical locations. Yet what unifies these individuals into a single agent (i.e., "Apple") capable of forming intentions and being held responsible for wrongful and praiseworthy actions is not their mere aggregation, but their participation in a shared corporate structure. This structure includes formal and informal hierarchies of decision-making, legally recognized norms, internal feedback mechanisms, and a coherent public-facing brand. On Hess's account, Apple, as an agent, is unified not simply by these individuals, but by their patterned participation in a complex, enduring structure that "embodies a rational point of view".[28] In other words, the corporate structure constitutes a unified standpoint—a logically integrated set of "commitments about fact and value"—from which the corporate agent can deliberate and act.[29] The structure thereby provides a coherent perspective for evaluating reasons, forming intentions, and guiding action at the collective level, even when no individual member fully shares or represents that perspective. Taken together, these considerations yield corporate agents as unified, materially anchored, and socially structured entities.

---

[26] Hess, 'Re-bunking Corporate Agency', 990–91.

[27] Hess, 'Re-bunking Corporate Agency', 988. The use of the term "autonomous" here might be seen as suggestive of emergence. To appreciate how far Hess develops this notion of autonomy: Björnsson, Gunnar, and Kendy Hess (2017) 'Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents', *Philosophy and Phenomenological Research* **94**: 273-98. Interestingly, in that paper, Björnsson and Hess acknowledge disagreement on precisely the issue at stake in the present paper: "We disagree among ourselves about how best to think about the relation between a corporate agent and its "members". Björnsson and Hess, 'Corporate Crocodile Tears?', 276 fn. 13.

[28] Hess, 'Re-bunking Corporate Agency', 984; Rovane, Carol (1998) *The Bounds of Agency: An Essay in Revisionary Metaphysics*, Princeton University Press.

[29] Hess, 'Re-bunking Corporate Agency', 985.

This paper aims to clarify an apparent tension in this story. On the one hand, Hess maintains that a corporate agent is nothing over and above its members; that is, the corporate agent *is* the members, and more.[30] On the other hand, she argues that the holist can maintain that corporate agency is "far more autonomous vis-à-vis corporate members than the other theories can account for".[31] It is not immediately clear how these two commitments sit together. Hence, how they are to be reconciled calls for further clarification. With Hess's holism in place, I now turn to the two metaphysical relations that will later be used to further explicate the relation between corporate agents and their members: grounding and emergence. The central contribution of this paper lies in articulating the nexus between Hess's holism and these metaphysical relations, which have not previously been integrated into her account.

1.2 Grounding

The notion of grounding has gained prominence in contemporary analytic metaphysics since the early 2000s. For the purposes of the present paper, grounding is understood as an asymmetric relation of non-causal metaphysical dependence between entities.[32] To say that one entity is grounded in another is to say that the former exists *in virtue of* the latter. Grounding thus serves two closely related functions. First, it structures reality in terms of relative fundamentality.[33] Second, it provides metaphysical explanation. Grounding tells us not merely that two entities are connected but why one depends upon the other. Another important feature of grounding, particularly for Hess's account of corporate agents, is that grounds are sufficient for the entities they ground (their "groundees"). If an entity genuinely grounds another, then the existence of the grounds is sufficient for the existence of the groundee. The existence of a statue, for example, is grounded in a particular arrangement of clay. Once the relevant arrangement obtains, the clay is sufficient for the existence of the statue. A further consequence of this non-causal character is that grounding is synchronic: unlike causal relations, it does not involve temporal succession between grounds and groundees but rather concerns the metaphysical dependence that holds between entities at a given time.

These considerations allow us to characterize non-reductive grounding in terms of four interrelated conditions. An entity x is non-reductively grounded in entities y_1, y_2, …, y_n iff:

[30] Hess, 'The Peculiar Unity of Corporate Agents'; Hess, 'Assembling the Elephant'; Hess, 'Re-bunking Corporate Agency'.
[31] Hess, 'Re-bunking Corporate Agency', 988.
[32] Unless otherwise indicated, "entities" in this paper refers to material entities.
[33] Fundamentality, as used here, tracks metaphysical priority. An entity is more fundamental than another when the latter exists in virtue of the former. More fundamental entities thus serve as ontological grounds, while less fundamental entities are grounded in them. Grounding may therefore be understood as a relation that structures reality according to relative levels of metaphysical dependence.

(i) x exists in virtue of $y_1, y_2, \ldots, y_n$, such that the latter provide the metaphysical basis for x;

(ii) $y_1, y_2, \ldots, y_n$ are metaphysically prior to x and explain why x exists;

(iii) x is not identical with $y_1, y_2, \ldots, y_n$; and

(iv) although dependent upon $y_1, y_2, \ldots, y_n$, x possesses a distinct ontological status and is not reducible to them.

I shall refer to these conditions as (i) "in-virtue-of dependence," (ii) "metaphysical priority," (iii) "non-identity," and (iv) "irreducibility".

My use of grounding in the present paper departs in two respects from many standard formulations. First, I employ entity-grounding rather than fact-grounding. Much of the grounding literature characterizes grounding as a relation between facts or truths.[34] By contrast, my concern is with the relation between entities themselves. The relevant question is not primarily whether facts about corporate agents are grounded in facts about members, but whether corporate agents as entities are grounded in their members. This reflects a commitment to a metaphysical framework in which explanatory priority is understood in terms of entities themselves, rather than in terms of facts about them. Accordingly, the relata of the grounding relation in this paper are entities.[35] Second, the grounding relation at issue is non-reductive. Non-reductive grounding is a relation of ontological dependence that does not entail that the groundee can be reductively explained in terms of its grounds. A non-reductively grounded entity may depend upon more fundamental entities for its existence while nevertheless possessing a nature that is not exhaustively captured by those entities alone.[36]

[34] Rosen, Gideon (2010) 'Metaphysical Dependence: Grounding and Reduction', in Hale, Bob and Aviv Hoffmann, eds., *Modality*, Oxford University Press; Fine, Kit (2012) 'Guide to Ground', in Fabrice Correia and Benjamin Schnieder, eds., *Metaphysical Grounding*, Cambridge University Press.

[35] For what it is for the existence of certain entities to depend on the existence of others (within a fact-only account of grounding): Schnieder, Benjamin (2020) "Grounding and Dependence," *Synthese* **197**: 95-124.

[36] Jonathan Schaffer and others often keep grounding flexible enough to allow non-reductive versions: Schaffer, Jonathan (2009) 'On What Grounds What', in Ryan Wasserman, David Manley, and David Chalmers, eds., *Metametaphysics: New Essays on the Foundations of Ontology*, Oxford University Press. For some remarks concerning non-reductive grounding: Rosen, Gideon (2010) 'Metaphysical Dependence: Grounding and Reduction', in Bob Hale and Aviv Hoffmann, eds., *Modality*, Oxford University Press. It should be noted, however, that whereas Rosen defends a fact-only account of grounding, my framework permits grounding relations in which entities, and not facts, may serve as grounds and groundees. This indicates that entity-grounding is not without controversy. For a further fact-only account of grounding: Fine, Kit (2012) 'Guide to Ground', in Fabrice Correia and Benjamin Schnieder, eds., *Metaphysical Grounding*, Cambridge University Press. Nevertheless, I take it that no extended defense of entity-grounding is required here, since the aim of this paper is not to resolve the fact-versus-entity-grounding debate, but to develop the consequences of an entity-grounding framework for the corporate agent–member relation. For a sustained argument in favor of entity-grounding over fact-only-grounding: Wilhelm, Isaac (2020) 'An Argument for Entity Grounding', *Analysis* **80**: 500-7.

Grounding should be distinguished from supervenience. The latter has been widely used across many areas of philosophy to analyze non-reductive dependence relations between different kinds of properties, most notably between mental and physical properties. Supervenience captures a pattern of covariation: any difference at the supervenient level requires a corresponding difference in the underlying base. Yet supervenience alone remains silent as to why such covariation obtains. It establishes dependence only in a modal sense, without explaining the metaphysical basis of that dependence. This limitation has motivated increasing interest in grounding. Unlike supervenience, grounding is typically understood as an asymmetric, explanatory relation that reveals what metaphysically accounts for what. If corporate agent-level facts were captured only by supervenience on member-level facts, we would know that no change in the corporate agent could occur without some corresponding change among members. We would not yet know what explains the corporate agent's existence. Since the present inquiry concerns metaphysical dependence rather than mere covariance, grounding provides the more appropriate framework. Section 2 develops this point further.

1.3 Emergence

This brings us to emergence. The term "emergence" has been used across philosophy and the sciences in many distinct ways. At its most general, emergence is invoked whenever a phenomenon appears to depend upon a more basic domain while nevertheless exhibiting a degree of autonomy from it. The challenge is to specify the nature of this autonomy.

A useful starting point is Chalmers's distinction between weak and strong emergence.[37] Weak emergence characterizes phenomena that arise from a lower-level domain and are in principle derivable from it, even if such derivation requires extensive computation or simulation. The autonomy involved is therefore epistemic rather than ontological: the phenomenon may be surprising or difficult to predict, but it remains wholly determined by and derivable from its base.[38] For instance, the intricate patterns formed by a flock of birds can, in principle, be derived from the interaction of simple behavioral rules governing individual birds, even if doing so requires extensive simulation. Strong emergence, by contrast, characterizes phenomena whose truths are not deducible even in principle from a complete specification of the lower-level domain.[39] Strong emergence is often characterized as involving metaphysical rather than merely epistemic autonomy. The paradigmatic example of strong emergence, and the primary focus of Chalmers's discussion, is consciousness: if conscious experience cannot, even in principle, be deduced from a

[37] Chalmers (2006) 'Strong and Weak Emergence', in Philip Clayton and Paul Davies, eds., *The Re-emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, Oxford University Press.
[38] Chalmers, 'Strong and Weak Emergence', 244.
[39] Chalmers, 'Strong and Weak Emergence', 244.

complete specification of all physical facts, then consciousness would be strongly emergent.[40]

While these notions have proven useful in debates concerning phenomena such as consciousness, neither fully captures the relation between the corporate agent's features and its members that is implicit in Hess's holism. Weak emergence is too weak because the autonomy of corporate agents' capacities and properties is not merely epistemic. The claim is not simply that these capacities and properties are difficult to predict or model. Rather, the corporate agent possesses capacities and properties that belong to it as a unified material entity. These features are not merely invoked as convenient descriptions of member-level activity. A merely weakly emergent corporate agent would remain exhaustively explicable through the behavior of its members. This would leave out the ontological claim Hess seeks to defend. Strong emergence, however, is too strong for Hess's purposes. Defending the claim that the corporate agent's capacities and properties are, in principle, not derivable from a complete specification of the corporate agent's members and their capacities and properties would require a substantially stronger metaphysical commitment than Hess's account appears to require. Hess instead emphasizes that corporate-level capacities arise from the internal mechanisms of the corporate agent, enacted through the participation of members occupying organizational roles. Moreover, she maintains that many capacities exist at both the individual and corporate levels, differing not in kind but in their mode of realization and scope.[41] Such a view does not require positing a fundamentally autonomous ontological domain beyond the members and their social relations. Put differently, commitment to strong emergence would sit uneasily with Hess's aim of re-membering the corporate agent, insofar as it would risk obscuring the dependence of corporate-level capacities on the members and social structure from which they emerge.[42]

For these reasons, I introduce a modified notion of emergence, which I shall call "emergence+". The "+" in emergence+ signals that the notion retains what is philosophically valuable in emergence—the idea that higher-level features represent a genuine addition to the metaphysical picture—while rejecting the stronger implication that such novelty requires a break in dependence relations. Emergence+ therefore concerns novelty within dependence rather than novelty through independence. A set of features *F* of an entity *x* is emergent+ from entities y_1, y_2, …, y_n, relative to the set of features *P* instantiated by those entities iff:

(i) *F* belongs to an entity *x* that depends upon an underlying basis consisting of y_1, y_2, …, y_n;

(ii) *F* depends upon the organization or arrangement of y_1, y_2, …, y_n;

(iii) *F* belongs to *x* as a whole rather than to y_1, y_2, …, y_n considered individually or as an aggregate; and

[40] Chalmers, 'Strong and Weak Emergence', 246–50.
[41] Hess, 'Re-bunking Corporate Agency', 994.
[42] Hess, 'Re-bunking Corporate Agency', 984.

(iv) *F* cannot be fully explained in terms of *P*.

Let us call these four conditions (i) "dependence," (ii) "organizational dependence," (iii) "whole-level possession," and (iv) "non-exhaustive explanation."

The distinctive feature of emergence+ lies in the combination of (i) dependence and (ii) organizational dependence: the relevant novelty does not arise from the entity's independence from its underlying basis, but from the way in which that basis is *organized*. Emergence+ therefore locates the source of emergent features neither in the isolated elements themselves nor in an autonomous higher-level domain, but in the structured relations among those elements. At the same time, the condition of non-exhaustive explanation preserves the anti-reductionist insight that such features cannot be fully captured by an account of the underlying basis alone. Emergence+ thus occupies an intermediate position between reductive explanation and stronger forms of emergence that posit a more radical discontinuity between levels.

§ 2. Grounding Corporate Agents

This section argues that corporate agents are grounded in their members. As outlined in the previous section, grounding, for the purposes of the present paper, is understood as an asymmetric relation of non-causal metaphysical dependence between entities that explains why one entity exists in virtue of another. My argument goes as follows:

P1. If an entity x satisfies conditions (i)–(iv) with respect to entities y_1, y_2, …, y_n, then x is non-reductively grounded in y_1, y_2, …, y_n.

P2. [(i) *in-virtue-of dependence*] The corporate agent exists in virtue of its members, whose structured participation provides the metaphysical basis for its existence.

P3. [(ii) *metaphysical priority*] The members of the corporate agent are metaphysically prior to the corporate agent and explain why it exists.

P4. [(iii) *non-identity*] The corporate agent is not identical with its members.

P5. [(iv) *irreducibility*] Although dependent upon its members, the corporate agent possesses irreducible properties and capacities that cannot be fully explained in terms of the properties and capacities of its members considered individually or as a mere aggregate.

C. Therefore, the corporate agent is non-reductively grounded in its members.

The argument above can be understood as an attempt to make explicit the metaphysical relation between corporate agents and their members that is implicit in

Hess's holist account.[43] Each premise corresponds to a central commitment of the account. First, Hess emphasizes that corporate agents do not exist independently of the individuals who participate in them. Rather, members are an integral part of the corporate agent's existence and function.[44] This justifies (P2): the corporate agent exists in virtue of its members, whose structured participation provides the metaphysical basis for its existence. Second, because the existence of the corporate agent depends upon the existence and participation of its members, Hess's account provides the basis for treating the members as metaphysically prior to the corporate agent, thereby establishing (P3).[45] Third, Hess's account supports the claim that the corporate agent is not identical with its members. Although the existence of a corporate agent depends upon the individuals who participate in it, Hess rejects the view that the corporate agent is simply the collection of those individuals. A corporate agent is a unified entity that embodies a rational point of view belonging to the corporate whole, rather than to any particular member or to the members considered merely as a plurality. [46] The members therefore ground the existence and function of the corporate agent without being identical with it. This justifies (P4). Finally, Hess claims that the corporate agent is irreducible to its members. She argues that the corporate agent possesses properties and capacities (e.g., the capacity to deliberate and act) that cannot be fully explained in terms of the properties and capacities of its members considered individually or as a mere collective.[47] This establishes (P5): although dependent upon its members, the corporate agent retains a distinct ontological status. Taken together, these premises show that Hess's account satisfies the conditions for non-reductive grounding.

The central intuition behind this argument is that corporate agents are neither independent entities detached from their members nor merely aggregates of those members. Non-reductive grounding captures this intermediate position: it explains the corporate agent's dependence on its members while preserving the claim that the corporate agent is not identical with, or reducible to, them. Hess's notion of "re-membering" the corporate agent is particularly important in this regard. Against views that risk separating corporate agents from the individuals who participate in them, Hess argues that members are essential to what corporate agents are. At the same time, she rejects the reductionist conclusion that the corporate agent is merely

[43] The grounding relation defended here should be distinguished from other metaphysical relations discussed in connection with corporate agents, such as composition and constitution. These relations address a different metaphysical question. Composition concerns the relation between parts and wholes, while constitution concerns a synchronic relation of necessity between two entities in which one entity constitutes another under certain conditions. The present proposal does not take a position on whether corporate agents should be understood in compositional or constitutional terms, since these debates concern the ontological status of corporate agents as objects rather than the dependence relation between corporate agents and their members that is the focus of this paper. More specifically, Hess's discussion of composition and constitution is primarily concerned with whether "the corporate agent qualifies as a material object in its own right". Hess, 'The Peculiar Unity of Corporate Agents', 36.

[44] Hess, 'Re-bunking Corporate Agency', 989.

[45] Hess, 'Re-bunking Corporate Agency', 988–9.

[46] Hess, "Re-bunking Corporate Agency', 988–9; 994–6.

[47] Hess, 'Re-bunking Corporate Agency', 983–5; 994–6.

its members taken collectively. The internal mechanisms of the corporate agent enable a kind of agency that is distinct from that of its members. This agency will be the focus of Section 3. The grounding relation articulated here captures Hess's dual commitment: corporate agents depend upon their members for their existence without being identical to or exhaustively explainable in terms of those members.

Consider, for example, Ströck-Brot GmbH (hereafter Ströck), a well-known Austrian bakery chain. Ströck exists as a corporate agent capable of entering contracts, employing workers, making strategic decisions, and pursuing long-term objectives. Yet none of these capacities could exist independently of the individuals who participate in the organization. Managers negotiate with suppliers, employees produce and distribute goods, and administrative staff maintain the practices and systems necessary for the organization's continued operation. In this sense, Ströck exists in virtue of its members: without those individuals participating in the relevant organizational roles and relations, there would be no corporate agent. However, this dependence does not entail that Ströck is identical with its members. Ströck is not simply the aggregate of the individuals who work for it at a given time. The organization persists despite changes in personnel, and the entity that enters contracts, owns property, and pursues organizational objectives is Ströck itself rather than any individual member or collection of members. The relevant dependence is therefore not one of identity but of grounding: the members provide the metaphysical basis for Ströck's existence as a corporate agent, while Ströck remains a distinct entity.

The grounding notion presented here should be distinguished from supervenience. One might be tempted to claim that the corporate agent merely supervenes on its members. Certainly, on Hess's account, there is a sense in which no change in a corporate agent can occur without some corresponding change among its members. If Ströck were to adopt a new pricing policy, modify its organizational hierarchy, or cease operations altogether, some change in the actions, intentions, or relations of its members would necessarily accompany that alteration. Supervenience successfully captures this pattern of covariance. Yet covariance alone does not provide metaphysical explanation. Supervenience tells us that there can be no difference at the corporate level without a difference at the member level, but it remains silent concerning why this dependence obtains. The supervenience thesis leaves open the explanatory question of what accounts for the corporate agent's existence as a unified entity in the first place. Grounding addresses precisely this question. To say that the corporate agent is grounded in its members is not merely to assert that member-level changes track corporate-level changes. It is to claim that the corporate agent exists in virtue of the structured participation of its members. Grounding therefore provides an explanatory framework that supervenience cannot provide.

Although non-reductive grounding captures the metaphysical profile of the corporate agent as Hess intends to characterize it, one might argue that my argument leaves unspecified how exactly the corporate agent comes to possess capacities and properties that are not exhausted by its members. While (P5) establishes the irreducibility of these features, the nature of this irreducibility requires further clarification. The notion of emergence+, to which I now turn, provides a framework for clarifying this.

§ 3. Emergent+ Capacities

This section argues that the corporate agent's capacities and properties are emergent+ from its members. Put differently, this section aims at clarifying (P5). My argument goes as follows:

> P1. If a set of features F satisfies conditions (i)–(iv) with respect to an entity x, an underlying basis consisting of entities $y_1, y_2, \ldots, y_n$, and the set of features P instantiated by those entities, then F is emergent+ from $y_1, y_2, \ldots, y_n$.
>
> P2. [(i) *dependence*] The corporate agent's capacities and properties belong to a corporate agent that is non-reductively grounded in its members.
>
> P3. [(ii) *organizational dependence*] The corporate agent's capacities and properties depend upon the organization of its members.
>
> P4. [(iii) *whole-level possession*] The corporate agent's capacities and properties belong to the corporate agent as a whole rather than to individual members or their aggregate.
>
> P5. [(iv) *non-exhaustive explanation*] The corporate agent's capacities and properties cannot be fully explained in terms of the capacities and properties of its members.
>
> C. Therefore, the corporate agent's capacities and properties are emergent+ from its members, relative to the capacities and properties instantiated by those members.

The argument above aims to show that the capacities and properties of corporate agents satisfy the four conditions for emergence+ introduced in Section 1.3. The first condition is secured by the grounding relation established in Section 2: because corporate agents are non-reductively grounded in their members, the capacities and properties they possess belong to an entity whose existence depends upon an underlying basis consisting of those members. This establishes (P2).

The second condition concerns the source of corporate-level capacities. Hess argues that corporate agency does not arise from the mere aggregation of individual agents, but from the organization of members within an enduring social structure that coordinates their activities and embodies a unified rational standpoint.[48] This organizational structure explains why corporate capacities depend upon the

[48] Hess, 'Re-bunking Corporate Agency', 985–9.

arrangement of members rather than upon the members considered independently, thereby establishing (P3).

The third condition follows from Hess's claim that the corporate agent embodies a unified rational point of view.[49] This standpoint enables the corporate agent to deliberate, evaluate reasons, form intentions, and act over time as a single entity. Importantly, this is entirely consistent with Hess's insistence that the unity of the corporate agent is realized through the organized participation of its members. The members collectively enable the standpoint, but they do not individually possess it. Once such a unified rational standpoint exists, capacities such as rational agency, deliberation, and intention-formation are attributable only to the corporate agent itself rather than to any individual member or to the members considered merely as an aggregate.[50] This establishes (P4).

Finally, and most importantly, Hess rejects the reductionist claim that corporate intentionality is simply identical with the intentionality of members. According to Hess, corporate intentionality differs from member intentionality in its "source, its duration, its inputs, and above all in its content and implications, among other things."[51] Consequently, Hess denies that the intentionality of a corporate agent is simply "numerically identical to the intentionality of the members, summed or otherwise".[52] This does not amount to claiming that the capacities of corporate agents cannot be *qualitatively* identical or similar to those of their members. The claim is only that they remain *quantitatively* distinct, existing on both levels. This establishes (P5): the corporate agent's capacities and properties cannot be fully explained in terms of the capacities and properties of its members. Taken together, these premises show that the corporate agent's capacities and properties satisfy the conditions for emergence+.

The central intuition behind this argument is that although corporate capacities depend entirely upon members, they do not belong to members. They arise only once individuals are organized into the kind of enduring social structure described by Hess. Emergence+ therefore captures both sides of Hess's position: dependence upon members and genuine whole-level agency. The point is not that corporate agents possess mysterious or independent powers, but that organization itself makes possible capacities that are attributable only to the corporate whole.

To appreciate Hess's point here, consider Apple. Apple formulates long-term strategic objectives and responds to regulatory challenges. On Hess's account, these actions are guided by a corporate point of view. Yet no individual employee fully occupies that point of view. Individual members each occupy only a partial perspective on the corporate agent's activities. Even the chief executive officer does not exhaustively represent Apple's rational standpoint. Nevertheless, Apple acts as though guided by a unified perspective that persists despite turnover among

[49] Hess, 'Re-bunking Corporate Agency', 984–5.
[50] Hess, 'Re-bunking Corporate Agency', 984–5.
[51] Hess, 'Re-bunking Corporate Agency', 996.
[52] Hess, 'Re-bunking Corporate Agency', 996.

personnel and disagreement among members. The existence of such a standpoint cannot be fully explained in terms of individual agency. Certainly, the standpoint depends upon the activities of individuals. Without members there would be no corporate point of view. Yet the standpoint itself cannot be identified with the perspective of any particular member or collection of members. It belongs to the organized whole and is thus only attributable to the corporate entity itself. This conclusion becomes even clearer when one considers the distinctive capacities exercised by corporate agents. Corporate agents can enter contracts, incur debts, establish organizational policies, acquire property, initiate litigation, and distribute dividends. Although these capacities are exercised through the actions of members, they are not capacities of those members themselves. Apple's decision to prioritize privacy over particular forms of data monetization, for example, is attributable to Apple as a corporate agent rather than to any individual employee or collection of employees. Numerous individuals contribute to the policy, yet no single member determines it or occupies the corporate standpoint from which it is adopted. The capacity therefore belongs to the organized whole, even though it depends entirely upon the coordinated participation of its members.

The argument developed in this section helps to clarify an important aspect of Hess's holism. Emergence+ provides a metaphysical framework for understanding how capacities can remain dependent upon members while nevertheless belonging only to the corporate agent. The members, organized through the corporate structure, make these capacities possible, but they do not thereby exhaustively explain them. This naturally raises the question of how emergence+ relates to the grounding relation defended in the previous section. The final task of this paper is therefore to show that grounding and emergence+ are complementary components of a unified metaphysical framework.

§ 4. Reconciling Grounding and Emergence+

At first glance, grounding and emergence+ appear to be in tension. Grounding emphasizes metaphysical dependence: corporate agents exist in virtue of their members. Emergence+, by contrast, emphasizes the novelty and irreducibility of certain corporate-level capacities. One might therefore suspect that the two notions compete to explain the same phenomenon or that emergence+ becomes explanatorily redundant once grounding has been established. I argue that neither concern is well founded. Grounding and emergence+ answer different metaphysical questions and therefore occupy complementary explanatory roles.

The key to reconciling the two notions lies in distinguishing the corporate agent from the capacities it possesses. Grounding concerns the existence of the corporate agent itself. It explains why there is such an entity at all and identifies the members, organized through an enduring social structure, as its metaphysical basis. Emergence+, by contrast, concerns the capacities and properties of that already grounded entity. It explains why the corporate agent possesses forms of intentionality, agency, and causal efficacy that cannot be exhaustively explained by reference to the capacities possessed by its members.

This distinction parallels one emphasized by Hess herself: the distinction between the corporate *agent* and corporate *agency*. As Hess notes, the corporate agent is "the single metaphysically robust entity that bears the capacity of agency," whereas corporate agency concerns the nature and exercise of that capacity.[53] She further observes that how we understand the corporate agent has significant implications for how we understand its agency.[54] The present proposal makes this relationship more precise. The corporate agent is non-reductively grounded in its members, whereas its capacities—including the capacity for agency—are emergent+ from those members.

Once this distinction is recognized, the charge of explanatory redundancy loses its force. Grounding and emergence+ do not explain the same aspect of corporate ontology. Grounding answers the existential question: what explains the existence of the corporate agent? Emergence+, by contrast, answers the characterization question: what explains the distinctive capacities possessed by that grounded entity? Because these are different explanatory projects, neither relation renders the other superfluous. Rather than introducing an additional explanatory layer alongside grounding, emergence+ makes explicit the anti-reductionist commitment that constitutes the grounding argument's fifth premise. Its role is therefore not to supplement grounding but to clarify the metaphysical status of certain higher-level features of an already grounded entity. The resulting picture captures both dimensions of Hess's holism. Corporate agents are thoroughly dependent upon their members for their existence, yet the organization of those members gives rise to capacities properly attributable only to the corporate whole. Grounding therefore secures dependence without reduction, while emergence+ secures irreducibility without independence. Together they provide a unified metaphysical framework that explains both what corporate agents are and how they are capable of acting.

§ 5. Conclusion

This paper has sought to clarify the metaphysical relation between corporate agents and their members under Hess's holist account of corporate agents. On Hess's view, corporate agents depend for their existence on the structured participation of their members, yet they also exhibit capacities and properties that are not reducible to those members, summed or otherwise. To articulate this dual character, I defended a two-part account. First, I argued that corporate agents are non-reductively grounded in their members. Grounding explains why corporate agents exist and identifies the ontological basis upon which they depend. Second, I argued that the corporate agent's capacities and properties are emergent+ from its members, thereby explaining the irreducibility expressed by (P5) in the grounding argument. Crucially,

[53] Hess, 'Assembling the Elephant', 113–4.
[54] Hess, 'Re-bunking Corporate Agency', 980–981 fn. 5

emergence+ captures the irreducibility of those capacities without severing the dependence of the corporate agent on its members. Finally, I argued that the apparent tension between grounding and emergence+ dissolves once it is recognized that the two relations answer different metaphysical questions. Grounding concerns existence; emergence+ concerns the nature of the already grounded entity. Rather than competing explanations, they are complementary dimensions of a single metaphysical picture. By integrating grounding and emergence+, the paper provides a more precise account of the corporate agent–member relation implicit in Hess's holism and thereby advances the broader project of developing a metaphysically robust account of corporate agents.[55]

[55] I would like to express my sincere gratitude to Severin Rohrer, Sofia Lieby, the organizers and attendees of *Social Ontology 2026* in Kraków (Poland), and three anonymous referees for *The Undergraduate Philosophy Journal of Australasia* for their invaluable and constructive feedback, which significantly improved this paper. I am also grateful to Kendy Hess for an immensely helpful conversation during *Social Ontology 2026*. Moreover, I am deeply indebted to Niels de Haan for introducing me to the contemporary debate on collective responsibility and for his generous guidance in response to my questions. Finally, I extend my thanks to the editorial board of *The Undergraduate Philosophy Journal of Australasia* for providing undergraduate students with such meaningful opportunities.

References

Björnsson, Gunnar, and Kendy Hess (2017) "Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents," *Philosophy and Phenomenological Research* 94(2): 273–98.

Chalmers, David J (2006) "Strong and Weak Emergence," in *The Re-emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, eds Philip Clayton and Paul Davies, 244–54. New York: Oxford University Press.

Fine, Kit (2012) "Guide to Ground," in *Metaphysical Grounding*, eds Fabrice Correia and Benjamin Schnieder, 37–80. Cambridge: Cambridge University Press.

French, Peter (1984) *Collective and Corporate Responsibility*. New York: Columbia University Press.

Hess, Kendy (2018a) "Does the Machine Need a Ghost? Corporate Agents as Nonconscious Kantian Moral Agents," *Journal of the American Philosophical Association* 4(1): 67–86.

Hess, Kendy (2018b) "The Peculiar Unity of Corporate Agents," in *Collectivity: Ontology, Ethics, and Social Justice*, eds Kendy Hess, Victoria Igneski, and Tracy Isaacs, 35–60. London: Rowman & Littlefield.

Hess, Kendy (2020) "Assembling the Elephant: Attending to the Metaphysics of Corporate Agents," in *The Routledge Handbook of Collective Responsibility*, eds Deborah Tollefsen and Saba Bazargan-Forward, 113–26. New York: Routledge.

Hess, Kendy (2025) "Re-bunking Corporate Agency," *Inquiry: An Interdisciplinary Journal of Philosophy* 68(3): 979–99.

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

Rosen, Gideon (2010) "Metaphysical Dependence: Grounding and Reduction," in *Modality*, eds Bob Hale and Aviv Hoffmann, 109–35. New York: Oxford University Press.

Rovane, Carol (1998) *The Bounds of Agency: An Essay in Revisionary Metaphysics*. Princeton: Princeton University Press.

Schaffer, Jonathan (2009) "On What Grounds What," in *Metametaphysics: New Essays on the Foundations of Ontology*, eds David Chalmers, David Manley, and Ryan Wasserman, 347–83. Oxford: Oxford University Press.

Schnieder, Benjamin (2020) "Grounding and Dependence," *Synthese* 197(1): 95–124.

Wilhelm, Isaac (2020) "An Argument for Entity Grounding," *Analysis* 80(3): 500–07.

Undergraduate Philosophy Journal of Australasia www.upja.online
Vol. 8, August 2026

Who Bears the Climate Burden? The Limits of Beneficiary Responsibility in Climate Justice

Enora Richard[1]

Australian National University

Abstract

Do presently living populations who have benefitted from excessive historical emission bear duties towards those suffering from climate-induced-hardship? The Beneficiary Pays Principle holds that they might. It mandates that agents who benefit from injustice can incur duties towards the victims of this injustice in virtue of benefiting. This paper examines the plausibility of the Beneficiary Pays Principle as a principle of climate justice. I consider the principle's application to the task of distributing the global cost of responding to climate change and aim to assess whether having benefitted from polluting activities bears on the responsibility one incurs to contribute to climate assistance. I do so by evaluating two prominent justifications for the Beneficiary Pays Principle. I conclude that neither justification succeeds in vindicating benefit-based duties of climate assistance. I find that the duties each principle defends prove no less applicable to affluent agents more generally. This conclusion calls into question the validity of the Beneficiary Pays Principle as a principle of climate justice. Although benefit-based duties have strong intuitive appeal, when we follow the reasoning offered by the two justifications assessed here, we can at best justify broad duties for all affluent agents to address climate harm.

[1] Enora is a recent graduate with a Bachelor of Philosophy, Politics and Economics from the University of Western Australia and an Honours in Philosophy from the Australian National University. She is particularly interested in questions of climate justice, having recently completed a thesis examining whether benefiting from polluting activities generates moral duties for agents to address global climate harm.

§ 0. Introduction

Do residents of wealthy countries such as Australia bear moral obligations toward those suffering from climate-induced harms given that their present economic prosperity, infrastructure, and quality of life are, in part, attributable to a history of high-polluting activities? Proponents of the Beneficiary Pays Principle (BPP) argue that they do. This principle holds that those who benefit from acts of injustice or wrongdoing ought, in some instances, to take on the burden to remediate the resulting harm. Although rich individuals in today's industrialised states may have played no role in the production of the historical emissions that have enriched them—and may even have opposed such activities had they been around to have their say—the BPP affirms that where they nonetheless continue to reap the benefits of these activities, this has a bearing on the moral duties they incur today. This paper will examine the validity of the BPP as a principle of climate justice. Many environmental justice advocates will feel the intuitive pull of benefit-based duties; if you benefit from harm, it seems natural to infer that you ought to incur duties to remediate it. It is thus important to examine the moral foundations of this idea. In an attempt to assess the validity of benefit-based duties of climate assistance, this piece will investigate the reasoning behind two prominent justifications for the BPP as applied to the question of how we ought to distribute the global burden to address climate harm. I will provide a negative thesis, demonstrating that the core philosophical justifications for this intuitively tempting principle fail to vindicate genuinely beneficiary-based duties of climate assistance.

Assigning duties to address the climate problem presents a challenge because many original polluters are no longer around to foot the bill for their contribution to climate-change. This leaves the question open on how to distribute the burden to remediate the harms associated with historical emissions. Should we do so merely with respect to people's wealth? Or ought we pay attention to the differential ways in which peoples' wealth came about, i.e., remain sensitive to the fact that some people were enriched by polluting activities to a greater extent than others? In practice, both approaches are likely to overlap with respect to whom they assign duties because much global wealth derives from polluting activities.[2] It is, however, theoretically interesting and perhaps motivationally useful, to correctly explain the basis for reasons to act. All wealthy agents may have duties to address the plight of the poor, but the BPP mandates that those who benefit from harmful activities have distinct *beneficiary-based* duties to remediate those who were harmed by such activities.

I distinguish between two distinct kinds of climate-induced harm. The first is the deprivation of present communities in less industrialised states of their capacity to generate energy free from the threat of detrimental climate change. Assuming the

[2] Singer, Peter (2002) *One World: The Ethics of Globalisation*, Yale University Press, 35; Page, Edward A (2012) 'Give It up for Climate Change: A Defence of the Beneficiary Pays Principle', *International Theory* **4**: 323.

global carbon budget[3] has been exhausted, any new emissions cannot be permitted if the international community is to remain within safe emission limits. Thus, where the global community chooses to set emission targets populations who have not yet used up their fair share[4] of the carbon budget are deprived of their capacity to do so. The second kind of climate harm is climate-inflicted hardship, such as extreme weather events or loss of homelands. My focus in this piece will be on this latter harm, what I refer to as *climate injury*. I engage with the normative question of how the global burden to address climate injury ought to be distributed among presently living populations. I seek to assess whether the BPP can plausibly guide this distribution—whether it can justify duties from those who benefit from over-emission to those presently harmed by climate change. I investigate two prominent justifications for the BPP as a means of assessing this principle. I conclude that both accounts fail to vindicate benefit-based duties of climate assistance.

After providing an exposition of the Beneficiary Pays Principle in Section 1, in Section 2, I examine a justification for this principle that grounds its normative force in the *attitudes* we ought to adopt towards wrongdoing. In Section 3, I then examine a justification for the BPP grounded in the value of *remediating* wrongdoing's effects. I argue that neither account successfully explains why those who have been enriched by historical emissions bear special duties to address climate harm that differ from the duties of other affluent agents. In demonstrating that neither foundational justification for the BPP succeeds in vindicating genuinely benefit-based duties of climate assistance, I call into question the BPP as a principle of climate justice. This is not to say that no version of the BPP will stand up to scrutiny but that, in the climate case, two primary ways in which this principle has at present been defended are unsuccessful.

§ 1. Distributing the Climate Burden

Duties to address the climate catastrophe must be assigned to present day agents and distributed justly. Three primary principles have been proposed to guide this distribution, with the BPP emerging as the most recent. To understand the role the BPP aims to play, we must first understand the limitations that other principles encounter. The first and most straightforward principle for distributing the climate burden is the Polluter Pays Principle (PPP). It holds that states ought to assume

---

[3] The carbon budget is the total net amount of carbon dioxide (CO2) that can still be emitted by human activities while limiting global warming to a specified level (e.g., 1.5°C or 2°C above pre-industrial levels) (IPCC, *Climate Change 2021: The Physical Science Basis* (2021) 2220). This budget is virtually exhausted if we are to remain within safe limits (Friedlingstein et al. (2025) 'Global Carbon Budget').

[4] There is debate about how the global carbon budget *ought* to be distributed among states and populations (Caney, Simon (2020) *Climate Justice*). We can assume here that there exists a means of conducting a fair distribution among states. Whether this be an equal division of the budget, a needs based distribution, or other, does not impact the conclusion that some states have exhausted their shares whilst others have not.

responsibilities for addressing the climate burden proportionally to their contribution to climate-change-inducing activities.[5] The most direct way in which we acquire moral obligations is as a result of the actions we perform.[6] With respect to distributing a collective burden, it is natural to think that those who have contributed the most to the problem ought to take on the largest burden to address it. This intuition is captured by the PPP. While this contribution-based approach is intuitively appealing in that it is backward-looking—it determines responsibility based on past actions—it encounters a difficulty: many of those who polluted in the past, and whose emissions contribute to present harm, are no longer around. If the PPP is to distribute the full burden to address all climate harm among living populations, it must conceive of states as temporally continuous agents. For current states to bear duties under the PPP to address climate harm proportionally to their state's *total* contribution to climate change, they must be regarded as the same agents that emitted in the past. This move is contentious from the standpoint of justice however, because the populations that constitute and govern states are not continuous over time. Present populations did not contribute to past emission production, and it may therefore be unfair to hold them responsible for these activities.[7] If we respect the moral independence of temporally distinct populations, polluter pays duties can only be assigned to living populations in proportion to their own share of total emissions. This would leave a large portion of climate harm—the proportion attributable to emissions that no one alive today contributed to—unaccounted for. Since almost half of emissions were produced pre-1990[8], a large proportion of emission activity will be attributable to people who are no longer around. Thus, if we seek to respect the moral independence of temporarily separate populations, whilst also distributing the full burden to address all climate harm among living populations, we need to appeal to another principle of climate justice, a principle that allows us to distribute among present populations a burden that they are not themselves responsible for.

As noted by Daniel Butt, it can be contentious to claim that agents incur moral obligations as a consequence of actions that others have performed.[9] Perhaps the most familiar cases in which agents acquire such duties are cases of rescue. If someone else pushes an innocent child into a pond and then runs away and you are now the only person around to rescue the child, you can acquire a duty to do so even though it was not you who pushed the child in.[10] Such duties are general and apply to anyone who has the capacity to help. The Ability to Pay Principle (APP) employs

---

[5] Neumayer, Eric (2000) 'In Defence of Historical Accountability for Greenhouse Gas Emissions', *Ecological Economics* **33**: 187-88.

[6] Butt, Daniel (2007) 'On Benefiting from Injustice', *Canadian Journal of Philosophy* **37**: 129.

[7] Caney, Simon (2010) 'Cosmopolitan Justice, Responsibility, and Global Climate Change' in *Climate Ethics : Essential Readings*, 130.

[8] Stainforth, Thorfinn and Bartosz Brzezinski (2020) 'More than Half of All CO2 Emissions since 1751 Emitted in the Last 30 Years', *IEEP AISBL*, https://ieep.eu/news/more-than-half-of-all-co2-emissions-since-1751-emitted-in-the-last-30-years/.

[9] Butt, 'On Benefiting from Injustice', 129.

[10] Adapted from: Singer, Peter (2016) *Famine, Affluence, and Morality*, Oxford University Press, 6.

this style of reasoning. It asserts that agents ought to contribute to collective endeavours in proportion to their ability to do so, and thus, that populations in high-income states ought to shoulder the majority of the climate burden merely owing to their greater comparative wealth.[11] The issue with such reasoning is that it is entirely forward looking, it no longer takes into account the unjust ways in which some wealth has come about. Wealth generated via historical enslavement is treated the same as wealth that derives from stumbling across a pot of gold. The BPP thus aims to introduce a way in which agents can acquire moral responsibility from actions that others perform in a manner that retains a backward-looking element. It identifies those who receive benefits from wrongful or unjust acts, as opposed to all those who are able, as having special remedial duties to compensate the victims of the resulting harm. In assigning duties on the basis of benefits received[12] rather than contributions made, it allows us to conceive of current populations as distinct from those who polluted in the past but to nevertheless remain sensitive to the ways in which emission generation has shaped present advantages.[13] In what follows I will provide an exposition of the BPP before moving into a more in-depth analysis of its justifications.

§ 1.1. The Beneficiary Pays Principle

The BPP is often formally understood as the following:

> 'If an agent A benefits from an action x that she did not perform, and an agent B is harmed by x, then, if B did not perform x, A, in virtue of benefiting from x, owes B compensation (which is not to exceed A's gain from x), insofar as no other agent has stronger duties to compensate B'.[14]

---

[11] Shue, Henry (1999) 'Global Environment and International Inequality', *International Affairs*, **75**, 537.
[12] There is contention about how a 'beneficiary' is to be defined. Beneficiary can be understood counterfactually, as someone who is 'better off than they otherwise would have been' (Butt, Daniel (2008) *Rectifying International Injustice: Principles of Compensation and Restitution between Nations*, Oxford University Press, 105–115). Or a beneficiary can merely be understood as someone who has 'received a benefit' (Barry, Christian and Robert Kirby, 'Scepticism about Beneficiary Pays: A Critique', *Journal of Applied Philosophy* **34**: 285-300). Here I take it that a beneficiary is an agent who has received a material or financial benefit from emission producing activities. A victim will be anyone who has experienced climate-related harm. It is not a problem that some agents have both received benefits and been harmed as a result of emission producing activities. I take it that there will be a global cost associated with remediating climate harm and that the question being asked is how ought we determine the responsibilities different groups have to contribute to this collective remediation fund. Thus, the question of who the specific victims are is not of fundamental importance here. Precisely how beneficiaries are defined also does not bear on the arguments in this piece. The question being asked is whether the notion of having received a benefit ought to bear, *in any respect*, on how much one ought to contribute to a global climate compensation fund, how one chooses to define beneficiaries and victims will not impact the structure of this investigation. Beneficiaries and victims can be understood as individuals, populations or states. This does not affect the arguments made here.
[13] Caney, 'Cosmopolitan Justice, Responsibility, and Global Climate Change', 127–28; Neumayer, 'In Defence of Historical Accountability for Greenhouse Gas Emissions', 189.
[14] Huseby, Robert (2015) 'Should the Beneficiaries Pay?', *Politics, Philosophy & Economics* **14**, 210.

The BPP defined in this way is designed to operate as a secondary principle. In cases where both Beneficiary Pays and Polluter Pays Principles apply, the former should not replace the latter.[15] In the climate case, if polluters are around and able, they will in the first instance be those who have a duty to remediate the victims of climate harm. The BPP applies where perpetrators cannot fulfil their duties, for example, where historical emitters are no longer around. It can also apply to beneficiaries who themselves caused the benefits and harms in question. For example, currently living individuals who pollute and also benefit from this. The focus here will be on *innocent* beneficiaries; those who benefit from emissions that they did not themselves contribute to. For example, present populations who have received benefits from historical emissions that were produced before they were around to have a say. Although many current agents who benefit from emission generating activities will also actively contribute to the production of new emissions (and will thus incur polluter-pays duties in virtue of this), the focus here is on whether they can derive duties in virtue of the benefits they have innocently received from past emissions. These historical emissions and their corresponding harm are of interest because they represent the gap that the Polluters Pays Principle fails to assign remediation duties for. No present agents can be assigned responsibility for these emissions on the PPPs reasoning. It is thus interesting to investigate whether the BPP can succeed in addressing this gap by assigning duties to innocent beneficiaries.

To see how the BPP is intended to operate, consider the following case that has been proposed in varying iterations by proponents of the BPP:

> *Missed interview*: Vic and Ben are both applying for the same job and Vic is clearly the best candidate. Peter, liking Ben and wanting him to get the job, sees an opportunity to ensure this by sabotaging Vic's car such that she cannot make it to the interview on time. Peter succeeds and as a result Ben is offered the job although he is less qualified than Vic. Peter then flees the scene and is nowhere to be found.[16]

Here, the BPP mandates that where Peter is not around to compensate Vic for the wrong he committed against her, Ben, in virtue of benefitting from Peter's wrongful act ought to take on this compensatory duty. The BPP assigns significance to the fact that certain benefits and harms have a 'common source'. It trades on the idea that there is something morally relevant about benefitting from the *same act* which causes another person harm. In other words, it is 'source sensitive'.[17] Ben is thought to bear a special corrective duty to Vic that he does not owe to someone else who suffers a comparable misfortune. Moreover, he is taken to have a special duty towards Vic that another agent who received a similar benefit to him, but not at Vic's expense, does not. This source sensitivity is intuitively compelling; there seems to be

[15] Huseby, 'Should the Beneficiaries Pay?', 211.

[16] Duus-Otterström, Göran (2017) 'Benefiting from Injustice and the Common-Source Problem', *Ethical Theory and Moral Practice* **20**, 1075; Page, Edward A and Göran Duus-Otterström, 'Correcting Unjust Enrichment: Explaining and Defending the Duty to Disgorge the Benefits of Wrongdoing', *Critical Review of International Social and Political Philosophy* **29**, 2–3.

[17] Duus-Otterström, 'Benefiting from Injustice and the Common-Source Problem', 1069.

something morally pertinent about the fact that Ben benefits because of an injustice committed against Vic that alters the character of his duties towards her from simply distributive ones to more targeted corrective ones. However, many object precisely to this feature of the BPP, contending that there is nothing morally relevant about benefits and harms sharing a common source. In what follows I will further explain source sensitivity and what might be troubling about it.

§ 1.2. Source Sensitivity

To provide a normative basis for source sensitive benefit-based duties, one must explain what is special about the cases in which these duties arise. People often benefit from actions that also cause harm to others in instances that don't seem to give rise to corrective duties. To see this, consider a case where corrective duties seem unreasonable.

> *Neighbours*: Two neighbours run businesses. Neighbour 1 has a tropical plant business and Neighbour 2, a solar energy business. Their area experiences a long period of cloudy rainstorms. Neighbour 1's tropical plants flourish as a result of this, and they manage to sell many more plants than they usually would. Neighbour 2 however does not get enough sun on their solar panels and thus struggles, making less profit than usual.

In this case, Neighbour 1 benefits from an event that harms Neighbour 2. However, this alone does not appear sufficient to ground a duty of compensation. Neighbour 1 may have distributive duties to assist Neighbour 2 by virtue of now being better off, but the mere fact that their benefit and Neighbour 2's harm share a common causal origin does not by itself establish a special moral duty toward Neighbour 2. The only connection between the neighbours is a causal one, and such a relation, on its own, seems too weak to generate directed duties of compensation. Where one, whether innocently or knowingly, performs an action that causes harm to another, this causal involvement may be relevant in grounding a duty to remediate that harm. However, being merely and innocently situated at the receiving end of a causal chain that also results in harm to another is insufficient to generate such a duty.[18] A justification for the BPP must give a "non-accidental" basis for source sensitivity and mere causal relation is insufficient.[19] We need an account of what distinguishes cases in which a causal relation gives rise to benefit-based duties from those in which it does not.

Proponents of the BPP typically provide this account by appealing to wrongdoing or injustice. On this view, the BPP applies in contexts where benefits and burdens arise from *wrongful* distortions—such as in Missed Interview—rather than from random natural events or mere brute luck—such as in Neighbours. Refining the BPP's scope to wrongful distortions does not, however, automatically save it from criticism.

[18] Huseby, 'Should the Beneficiaries Pay?', 219.
[19] Duus-Otterström, 'Benefiting from Injustice and the Common-Source Problem', 1069–71.

Consider the following case concerning four individuals whom we will call Wrongdoer, Victim, Beneficiary and Unfortunate:

> *Unfortunate:* In her dying moment, Wrongdoer unjustly harms Victim and, as a result of this, Beneficiary innocently benefits. When Victim is harmed, she becomes as badly off as Unfortunate, who suffers from some natural event.[20]

In this case, Beneficiary benefits from *wrongdoing*. If the BPP holds that it is normatively significant when benefits and harms flow from the *same wrongful act*, it entails that Beneficiary has a reason to address Victim's loss that she does not have to address Unfortunate's.[21] This, however, seems unfair. For critics of the BPP, the mere fact that Beneficiary and Victim are causally connected by Wrongdoer's unjust action is a superficial reason for Beneficiary to favour Victim over Unfortunate.[22]

In response to concerns about this unfairness the BPP may licence between the victims of different sorts of hardship, a proponent of the BPP can respond that this principle is not being advanced as the sole principle of justice. There are other principles that aim to equalize the effects of brute luck. In the *Unfortunate* example, Beneficiary may have *distributive* duties to Unfortunate. Where Beneficiary is the only agent capable of helping Victim or Unfortunate, they must simply weigh the demands that different principles make on them. This deliberation may give Beneficiary reason to prioritise Unfortunate, to prioritise Victim, or to distribute their benefits between them. The more fundamental question, however, is whether the BPP can successfully ground any *pro tanto* special duty owed by beneficiaries to victims. Critics such as Knight and Huseby contend not merely that source sensitivity risks licensing objectionable unfairness, but that its normative foundation is altogether unjustified. It is this question—whether *pro tanto* source-sensitive duties of assistance can be justified—that the present paper addresses.

The BPP's source sensitivity comprises two distinct claims. The first, *wrongdoing-or-injustice sensitivity*[23], requires that benefits generated by acts of wrongdoing or injustice be treated differently to wealth generated by good luck. The second, *directedness*, requires that benefits generated via unjust or wrongful acts be utilised to address the harms resulting from those same unjust acts, i.e., that beneficiaries' remedial duties ought to be 'directed'.[24] The latter is what makes the BPP unique. The notion that there is something wrong about holding benefits arising from wrongdoing or injustice can at most justify duties of disgorgement—duties to relinquish unjustly acquired wealth. The BPP holds that unjustly generated benefits

---

[20] Duus-Otterström, 'Benefiting from Injustice and the Common-Source Problem', 1069; Knight, Carl (2013) 'Benefiting from Injustice and Brute Luck', *Social Theory and Practice* **39**: 581–98.

[21] Knight, 'Benefiting from Injustice and Brute Luck', 588.

[22] Huseby, 'Should the Beneficiaries Pay?,' 218–20; Knight, 'Benefiting from Injustice and Brute Luck', 595–96; Lippert-Rasmussen, Kasper 'Affirmative Action, Historical Injustice, and the Concept of Beneficiaries', *Journal of Political Philosophy* **25**: 78–9.

[23] Term taken (and adapted) from: Duus-Otterström, Göran (2024) 'Climate Change, Historical Emissions, and Unjust Benefits: A Comment on Axel Gosseries' Account of Climate Justice', *Journal of Practical Ethics* **12**: 10–11.

[24] Duus-Otterström, 'Climate Change, Historical Emissions, and Unjust Benefits', 10–11.

ought not merely be disgorged but transferred to the victims of the injustice from which those benefits derive. Here, I assess the defensibility of directed duties owed by those who benefit from over-emission to those harmed by it. To justify directed beneficiary duties to address climate injury, a defence of the BPP must identify what is distinctive about emission beneficiaries that generates special duties to compensate the victims of climate harm which are not reducible to the general duties all affluent agents bear in virtue of their financial capacity to provide aid. In what follows I will present two prominent justifications of the BPP which hold that wrongdoing possesses a distinctive moral urgency—it ought to be avoided, condemned or corrected—and that this explains the directed duties wrongfully enriched emission beneficiaries incur towards the victims of climate injury. In Section 2, I present *attitudinal* defences of the BPP, which ground their normative force in the attitudes we ought to display towards wrongdoing, and in Section 3, I examine accounts that derive directed benefit-based duties from the moral value of *remediating* wrongful transactions. I argue that both rationales are ultimately unsuccessful in vindicating benefit-based duties of climate compensation.

§ 2. The Attitudinal Account

The notion that we ought to feel troubled when benefitting from wrongdoing is compelling. An act that is wrongfully committed is normatively distinct from a natural event; there is something about the former that is condemnable or worthy of disapproving attitudes. Butt's Attitudinal Inconsistency[25] justification for the BPP hinges on this intuition to vindicate directed duties. Butt asserts that we ought to display an aversion toward wrongdoing and that this requires a commitment to reversing its effects.[26] In the climate case, the fact that over-emitting is wrong means it is worthy of condemnation. The requirement to condemn generates corrective duties on the part of those who benefit to take actions in their power to remediate the resulting climate harm.

A first concern to address before I formally present Butt's account is that justifications for the BPP that trade on *wrongdoing* will only be applicable to those enriched by emissions that can be characterised as wrongful. Thus, such accounts may only apply to the beneficiaries of emissions produced after 1990, since prior to this point emitters could be excused on grounds of ignorance.[27] Given that approximately half of global carbon emissions have occurred since 1990[28], if any wrongdoing-sensitive versions of the BPP are defensible, they may still offer useful a

[25] Term taken from Page, Edward A and Göran Duus-Otterström (2023) 'M', *Critical Review of International Social and Political Philosophy*, 4.
[26] Butt, 'On Benefiting from Injustice', 143; Butt, Daniel (2014) 'A Doctrine Quite New and Altogether Untenable: Defending the Beneficiary Pays Principle', *Journal of Applied Philosophy* **31**: 340.
[27] Page, Edward A (2008) 'Distributing the Burdens of Climate Change', *Environmental Politics* **17**: 560; Page, 'Give It up for Climate Change', 313; Caney, Simon (2009) 'Justice and the Distribution of Greenhouse Gas Emissions 1', *Journal of Global Ethics* **5**: 134.
[28] Stainforth and Brzezinski, "More than Half of All CO2 Emissions since 1751 Emitted in the Last 30 Years."

basis for assigning duties to address a significant portion of the climate burden. A further contention is that some deny that *any* emissions can be characterised as wrongful or harmful. The dominant view among such critics is that no single or disaggregated act of emission makes a significant difference to the rate or extent of climate change and is therefore neither necessary nor sufficient for harm.[29] One reply to this causal critique is that all emissions can be considered harmful when harm is understood in expected rather than actual terms. Since climate change is harmful and emissions causally contribute to it, each act of emission carries a non-zero expectation of harm.[30] Alternatively, emissions might be considered wrongful regardless of whether they are understood to cause harm. Imposing a *risk* of harm on others via emitting greenhouse gasses may itself be taken to constitute a wrong.[31] I will not defend any particular account of what renders a population's emissions wrongful or harmful, but proceed on the assumption that one of these arguments succeeds in order to assess the plausibility of wrongdoing sensitive justifications of BPP in the climate case.

Butt's argument for the BPP rests on the following premise:

> "Taking our nature as moral agents seriously requires not only that we be willing not to commit acts of injustice ourselves, but that we hold a genuine aversion to injustice and its lasting effects. We make a conceptual error if we condemn a given action as unjust but are not willing to reverse or mitigate its effects on the grounds that it has benefited us. The refusal undermines the condemnation".[32]

This premise contains two important claims. The first is a moral one: we ought to demonstrate an aversion to wrongdoing. The second is a conceptual one: "there is an inconsistency in our moral outlook" if we condemn actions as wrong but then fail to take actions in our power to undo their effects.[33] There is a requirement to reverse the

[29] Sinnott-Armstrong, Walter (2010) 'It's Not My Fault: Global Warming and Individual Moral Obligations', in Walter Sinnott-Armstrong *Climate Ethics*, Oxford University Press.
[30] Broome, John (2019) 'Against Denialism', *The Monist* **102**: 110–29; Hiller, Avram (2011) "Climate Change and Individual Responsibility', *The Monist* **94**: 349–68.
[31] Barry, Christian and Garrett Cullity (2022) 'Offsetting and Risk Imposition', *Ethics* **132**: 352–81.
[32] Butt, 'On Benefiting from Injustice', 143.
[33] Butt, 'A Doctrine Quite New and Altogether Untenable', 340.

effects of wrongdoing in order to preserve our moral integrity.[34] This generates a justification for the BPP. Where agents are enriched by wrongdoing at another's expense, a moral requirement falls upon them to condemn this wrong. This requirement entails a demand to undo this wrongdoing's effects—the benefit and the harm—and thus requires that beneficiaries disgorge their benefit towards compensating victims. Where one is enriched by wrongful over-emission, the moral requirement that falls upon them to condemn these activities entails a demand that they utilise these benefits to redress climate-induced-harm. Failure to do so makes them "complicitous in, and or at least, not sufficiently opposed to, the wrongdoing".[35] In what follows, I present my critique of Butt's account.

§ 2.1. The Attitudinal Account and the Non-Distinctiveness Objection

Butt's argument fails to identify beneficiaries as agents who bear special duties to address the plight of those harmed by the unjust acts that enriched them. If, as Butt affirms, everyone ought to have an aversion to wrongdoing, this cannot provide beneficiaries with reason, over and above all others, to take actions in their power to redress the harm from which they benefit. In the climate injury case, the existence of a general duty to condemn wrongdoing, implies that all those who are able ought to act to undo the effects of wrongful over-emission. If one's duty to compensate a victim of wrongdoing derives from the value of upholding their moral integrity, or from a requirement to express concern for the victim's situation, then it seems any agent who recognizes the wrongfulness of harm towards a victim ought to take actions within their power to remediate it. It remains unclear what moral significance attaches to being the beneficiary of a wrongful act that renders one more blameworthy than any other competent and aware moral agent where they fail to act. It seems all well-off agents express the wrong sorts of moral attitudes where they choose not to address harm that it is in their power to remediate.

[34] Page and Duus-Otterström criticise this account for relying on the wrong sorts of moral reasons to motivate directed disgorgement—misdiagnosing what is morally significant about benefitting from injustice in focusing on restoring the beneficiary's moral integrity rather than addressing the injustice committed against the victim (Edward, A. Page and Göran Duus-Otterström, "Correcting Unjust Enrichment: explaining and defending the duty to disgorge the benefits of wrongdoing,", *Critical Review of International Social and Political Philosophy* (2023), 5). Similar reasoning could be employed however that shifts the focus away from the beneficiary's moral integrity towards their respect for the victim. Some suggest that failure to surrender benefits generated at a victim's expense amounts to *condoning* wrongdoing or harmdoing (Page, "Give It up for Climate Change," (2012), 315). As captured by Heyward, it might indicate that the wrongs committed against the victims "don't matter" and express "a morally repugnant attitude" towards the victims of injustice (Clare Heyward, "Benefiting from Climate Geoengineering and Corresponding Remedial Duties: The Case of Unforeseeable Harms," *Journal of Applied Philosophy* 31, no. 4, (2014), 410). This sort of reasoning similarly appeals to the moral attitudes we ought to have to ground duties for beneficiaries to relinquish wrongful gains. My critique of both these accounts is not that there's something unsatisfying about their style of reasoning, but that their justifications, by their own terms, cannot defend directedness in the climate injury case.

[35] Butt, 'A Doctrine Quite New and Altogether Untenable', 340.

What is special about beneficiaries is that they have *benefitted*. Wrongfully enriching actions have a *benefit-effect* (the endowment of benefits on one agent) and a *harm-effect* (the imposition of harm on another). On Butt's logic, those who are wrongfully enriched may plausibly have a duty to relinquish their benefits to undo the *benefit-effect* of the wrongful transaction that enriched them, but the *benefit-effect* is the only effect they are *specially* placed to undo. I contend, so long as beneficiaries disgorge their benefits, they have exhausted their special beneficiary-based duties. This special disgorgement duty could be adequately fulfilled where they use their benefits to compensate any victims of any unrelated harms, seeing as on Butt's account there is a general requirement to condemn all wrongful harm. For example, Australians who benefit from mining royalties could choose to donate a commensurate amount to a charity proving aid in conflict zones. The duty to compensate the victims of climate injury in particular will fall on all well-off agents, not merely those who've received emission benefits. Emission beneficiaries may have a duty to undo the *harm-effect* of over emission and thus to remediate the victims of climate harm, but they will incur this duty merely in virtue of being a moral agent who ought to be averse to such wrongdoing and is capable of helping victims. Where climate injury needs to be addressed, all well-off agents who are aware of this harm and fail to take actions in their power to remediate it, fail to express the appropriate attitudes towards these victims. Attitudinal accounts cannot explain why emission beneficiaries *in particular* manifest a morally objectionable attitude towards the victims of climate injury when they fail to contribute to climate compensation. It falls prey to what I term, the *non-distinctiveness objection,* the challenge that arises when a rationale offered for assigning duties to beneficiaries equally applies to all well-off agents, thereby undermining the claim that such duties are genuinely benefit-based.

To further assess the BPP's potential to generate duties in the climate case, I now turn to Edward Page and Göran Duus-Otterström's justification of the BPP. This account is grounded, not merely in the attitudes we ought to display towards wrongdoing, but in the independent value of correcting wrongful transactions. Although seemingly more robust, I will demonstrate that this account also falls vulnerable to the non-distinctiveness objection.

§ 3. The Corrective Justice Account

Page and Duus-Otterström present what they term the Corrective Justice Account (CJA) which grounds beneficiary-based duties in the "objective of correcting wrongful transactions".[36] Going back to Missed Interview, Page and Duus-Otterström argue that this situation needs to be remediated as a matter of corrective justice. Their rationale for directed duties arises from the aim of restoring the world to the state it would have been in had the wrongful transaction that

[36] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 11.

generated benefits and losses not occurred.[37] The duty Ben is supposed to incur to relinquish his benefits to Vic is one of corrective justice to remove the effects of Peter's wrongdoing from the world. Page and Duus-Otterström's argument for *directed* duties is that compared to inaction or undirected disgorgement, directed disgorgement more closely returns the world to a condition in which the wrongful act exerts no further influence over the prospects of any of the affected parties.[38] Where a wrongful transaction generates benefits and burdens, fully remediating the wrong requires redressing both the benefit-effect and the harm-effect. A directed transfer from the beneficiaries of an act to its victims brings both the victims and beneficiaries closer to the state they would have been in had the injustice not occurred and thus, "roundly" defeats the wrongdoing.[39] To undo Peter's wrongdoing wholly, Ben must compensate Vic. Simply relinquishing his benefit is insufficient to fully undo the effects of Peter's wrongdoing. I argue, contrarily, that wrongdoing can be roundly defeated without directed disgorgement. Directedness may be convenient—undoing both effects of a wrong in one transaction—but it is not *morally required*. I present my argument for this conclusion in what follows.

§ 3.1. The Corrective Justice Account and the Non-Distinctiveness Objection

According to the CJA, a beneficiary's duty to compensate the victims of an act that enriched them arises from their duty to remove the influence of a wrongful transaction from the world. I argue that beneficiaries are not specially placed to achieve this outcome. For Page and Duus-Otterström, although the value of undoing the effects of wrongdoing is impersonal and all agents incur duties to do so—an innocent beneficiary's reason to correct wrongdoing being no different to anyone else's—a beneficiary's receipt of benefits means they are *specially placed* to act.[40] As aforementioned, wrongful acts that generate benefits for one via harm to another have both a harm-effect and a benefit-effect. I argue that, on the CJA's reasoning, the only effect of a wrongful transaction a beneficiary is *specially* placed to undo is the benefit-effect.

If as Page and Duus-Otterström assert, our reasons to remove wrongdoing from the world are "general and the same for everyone"[41], it appears anyone who has the capacity to undo wrongful losses (the harm-effect) will incur a duty to do so. Cases presented by proponents of the BPP—such as Missed Interview—are often isolated such that the beneficiaries and victims of a singular unjust act are the only relevant agents in the picture.[42] Where this condition is not met however, the case for directed duties is less compelling. Where a wrongful act enriches an agent and harms another,

[37] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 11.
[38] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 12.
[39] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 12.
[40] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 11.
[41] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 11.
[42] Huseby, 'Should the Beneficiaries Pay?', 219.

if other well-off agents are around and aware of the wrongful harm to the victim, they will have the same duty as the beneficiary to undo its effects.

With respect to the benefit-effect on the other hand, because the beneficiary holds the benefit, they are specially placed to remove it from the world. The responsibility to undo this effect may generate a duty on the part of beneficiaries to relinquish their wrongfully acquired gain. However, where there are many poorly off agents present, and a beneficiary has a duty to undo the benefit-effect of the wrong that enriched them, they can do so by relinquishing their wrongful gains in the service of compensating any badly off agent. Where there are a range of wrongful harms occurring—such as war-inflicted deprivation, colonial injustice or, gendered oppression—emission beneficiaries will have equally pressing duties to use their unjustly acquired benefits to address these harms. On the condition that they forfeit their wrongful gains, their benefit-based duties are exhausted. So long as both the harm-effect and benefit-effect of a wrong are undone, the CJA's requirements have been fulfilled. Since this does not require a directed transfer from the beneficiaries of an act to its victims, Page and Duus-Otterström's account cannot provide a justification for directed duties of climate compensation.

Before I continue, I note that directed disgorgement might be morally required in one special type of case; cases where a beneficiary holds an object or quantum of value to which a victim is entitled, i.e., cases of restitution. Conveniently for Page and Duus-Otterström, Missed Interview, similarly to many cases employed by proponents of the BPP, can be articulated in terms of restitution. Ben received the job that Vic was entitled to and thus, may have duties of *restitution* to return it to her. The case of climate injury is not this type of case, however. Those who benefit from historical emissions don't hold what has been taken from those suffering from extreme weather events and other climate-change-induced hardships. Thus, I maintain that directed duties cannot be justified in this case.

Page and Duus-Otterström explain that one's receipt of a wrongful benefit activates their duty to take corrective action towards the victims of the wrong that enriched them.[43] I contend that what this activates however, is both a duty to disgorge their unjust benefit and an *ability* to compensate the victim. Where receiving a wrongful benefit engenders a beneficiary's ability to compensate the victim of the wrong that enriched them, it is tempting to infer that their duty towards this victim is benefit-based. But if, as per the CJA's reasoning, duties to compensate the victims of wrongs arise from the general goal of remediating wrongdoing, anyone who is able will incur a duty to act and thus, such duties are ultimately ability-based. Beneficiaries will have a duty to relinquish their unjust gains and a separate duty, in virtue of their new-found ability, to undo the wrong that the victim was subject to. The former will be benefit-based but the latter, ability-based. Page and Duus-Otterström assert that the CJA "enables us to recognize the intuitive pull of directedness without deriving [it] from the merely causal relationship shared by

[43] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 11.

beneficiaries and victims of a common injustice",[44] but I argue that they do exactly this. It is convenient and tempting to conflate the duty to disgorge and the duty to compensate. Where a beneficiary has been wrongfully enriched at a victim's expense they have on one hand, a duty to disgorge their unjust gains, and on the other, an ability-based duty to aid the victim. In such a case, a directed transfer from the beneficiary to the victim would be a straightforward way of roundly remediating an instance of wrongdoing; it would undo the benefit and harm effects of a wrong in one transaction. Such a directed transfer, is not morally required, however.

Where an isolated injustice occurs and there exists only one wrong that needs correcting, one agent who holds unjust gains and one agent that is unjustly harmed, the comprehensive correction of wrongdoing in this world will require a directed transfer from beneficiary to victim. However, the duties the enriched will have towards the victim will not arise in virtue of their benefitting from the same injustice that harmed them, rather they will arise because they have the capacity to help. The directedness required by isolated cases is merely accidental. So long as both effects of a wrong are undone, the CJA's requirements have been fulfilled and, in a world where multiple distortions are occurring, removing the effects of wrongdoing will not require directed transfers. Where injury to climate victims remains unaddressed, anyone who has the ability to act and is failing to do so wrongfully allows the effects of wrongdoing to persist. Those who hold emission benefits are not more culpable for failing to address this harm than others. Thus, the CJA falls prey to the non-distinctiveness objection and directed benefit-based duties of climate assistance cannot be mandated on its logic.

In a final attempt to defend directed duties of climate assistance, one might point out that wrongfully enriched emission beneficiaries have *weightier* reasons to relinquish their gains than others because they are unentitled to their wealth. One might claim that they have stronger duties to compensate the victims of climate harm than other well-off agents who have received their wealth justly. In response to this, I would purport that in the contemporary world, it is likely that most wealth will be the product of wrongdoing or injustice, and consequently unjustly held. Furthermore, we might think that so long as there exist very poorly off individuals who require extra wealth for subsistence, *any* excess wealth is unjustly held regardless of how it was generated.[45] If being unentitled to one's wealth is what gives an agent a strong duty to compensate the victims of wrongful harm, then many of today's affluent populations will incur equally pressing duties to relinquish their wealth towards funding a response to climate injury. The duties that the CJA mandates from the beneficiaries of unjust emissions to the victims of climate harm can be equally applied to all affluent agents. This account's normative basis might rest on the laudable goal of redressing the effects of wrongdoing but it cannot justify directed duties of climate assistance.

---

[44] Page and Duus-Otterström, 'Correcting Unjust Enrichment', 12.
[45] Robeyns, Ingrid (2022) 'Why Limitarianism?', *Journal of Political Philosophy* **30**: 249–70.

§ 4. Conclusion

The rationales for the BPP assessed in this paper were unable to successfully identify why the beneficiaries of historical emissions bear special duties, beyond those of other affluent agents, to compensate the victims of climate-induced harm. Unlike many critics of the BPP, I have not argued that the accounts I assessed employ flawed reasoning or lead to unacceptable conclusions. Rather, I have sought to demonstrate that where we take their premises and assumptions as acceptable, their normative foundations cannot, by their own lights, vindicate the conclusion that the beneficiaries of unjust emissions bear unique duties to the victims of climate harm. Both justifications fall prey to the non-distinctiveness objection. When following their reasoning, the duties they invoke fall on all agents who have the capacity act. Although the global population of emission beneficiaries is in practice likely to overlap quite substantially with the global affluent, and thus, duties to condemn or correct climate injustice will fall on this population regardless, these duties will not be distinctly benefit-based. This is not to say there is no way of defending the BPP's application to distributing the burden to address climate injury, but merely that the two foundational justifications for the principle assessed here, fail to explain what makes beneficiaries' moral position unique with respect to addressing climate harm.

References

Barry, Christian, and Garrett Cullity (2022) 'Offsetting and Risk Imposition', *Ethics* **132**: 352–81. DOI: 10.1086/716871

Barry, Christian, and Robert Kirby (2017) 'Scepticism about Beneficiary Pays: A Critique', *Journal of Applied Philosophy* **34**: 285–300. DOI: 10.1111/japp.12160

Broome, John (2019) 'Against Denialism', *The Monist* **102**: 110–29. DOI: 10.1093/monist/ony024

Butt, Daniel (2014) "'A Doctrine Quite New and Altogether Untenable': Defending the Beneficiary Pays Principle." *Journal of Applied Philosophy* 31: 336–48. DOI: 10.1111/japp.12073

Butt, Daniel (2007) 'On Benefiting from Injustice', *Canadian Journal of Philosophy* **37**: 129–52. Cambridge: Cambridge University Press.

Caney, Simon. "Cosmopolitan Justice, Responsibility, and Global Climate Change." *Climate Ethics : Essential Readings*, 2005 2010.

Caney, Simon. 'Justice and the Distribution of Greenhouse Gas Emissions', *Journal of Global Ethics* **5** (2009): 125–46. DOI: 10.1080/17449620903110300

Caney, Simon. 'Climate Justice', (2020). https://plato.stanford.edu/entries/justice-climate

Duus-Otterström, Göran (2017) 'Benefiting from Injustice and the Common-Source Problem', *Ethical Theory and Moral Practice* **20**: 1067–81. DOI: 10.1007/s10677-017-9845-7

Duus-Otterström, Göran (2024) 'Climate Change, Historical Emissions, and Unjust Benefits: A Comment on Axel Gosseries' Account of Climate Justice', *Journal of Practical Ethics* **12**: 6-15. DOI: 10.3998/jpe.6268

Friedlingstein, Pierre, Michael O'Sullivan, Matthew W. Jones, Robbie M. Andrew, Dorothee C. E. Bakker, Judith Hauck, Peter Landschützer, et al. 'Global Carbon Budget 2025', *Earth System Science Data* **18** (May 13, 2026): 3211–88. DOI: 10.5194/essd-18-3211-2026

Hiller, Avram (2011) 'Climate Change and Individual Responsibility', *The Monist* **94**: 349–68.

Huseby, Robert (2015) 'Should the Beneficiaries Pay?', *Politics, Philosophy & Economics* **14**: 209–25. DOI: 10.1177/1470594X13506366

Intergovernmental Panel on Climate Change (2021) 'Climate Change 2021: The Physical Science Basis', https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_FullReport.pdf

Knight, Carl (2013) 'Benefiting from Injustice and Brute Luck', *Social Theory and Practice* **39**: 581–98. DOI: 10.5840/soctheorpract201339433

Lippert‑Rasmussen, Kasper (2017) 'Affirmative Action, Historical Injustice, and the Concept of Beneficiaries', *Journal of Political Philosophy* **25**: 72–90. DOI: 10.1111/jopp.12092

Neumayer, Eric (2000) 'In Defence of Historical Accountability for Greenhouse Gas Emissions', *Ecological Economics* **33**: 185–92. DOI: 10.1016/S0921-8009(00)00135-X

Page, Edward A (2008) 'Distributing the Burdens of Climate Change', *Environmental Politics* **17**: 556–75. DOI: 10.1080/09644010802193419

Page, Edward A (2012) 'Give It up for Climate Change: A Defence of the Beneficiary Pays Principle', *International Theory* **4**: 300–330. DOI: 10.1017/S175297191200005X

Page, Edward A., and Göran Duus-Otterström (2023) 'Correcting unjust enrichment: explaining and defending the duty to disgorge the benefits of wrongdoing', *Critical Review of International Social and Political Philosophy* **29**: 693-718. DOI: 10.1080/13698230.2023.2245651

Robeyns, Ingrid (2022) 'Why Limitarianism?', *Journal of Political Philosophy* **30**: 249–70. DOI: 10.1111/jopp.12275

Shue, Henry (1999) 'Global Environment and International Inequality', *International Affairs (Royal Institute of International Affairs 1944-)* **75**: 531–45.

Singer, Peter. *Famine, Affluence, and Morality*. Oxford: Oxford University Press, 2016.

Singer, Peter. *One World: The Ethics of Globalisation*. With Brissenden collection and Australian National University Library Flood replacement project. Text Pub., 2002.

Sinnott-Armstrong, Walter. 'It's Not My Fault: Global Warming and Individual Moral Obligations', In *Climate Ethics*, by Walter Sinnott-Armstrong. Oxford University Press, 2010. DOI: 10.1093/oso/9780195399622.003.0029

Stainforth, Thorfinn, and Bartosz Brzezinski. "More than Half of All CO2 Emissions since 1751 Emitted in the Last 30 Years." *IEEP AISBL*, 2020.

https://ieep.eu/news/more-than-half-of-all-co2-emissions-since-1751-emitted-in-the-last-30-years/

Undergraduate Philosophy Journal of Australasia www.upja.online
Vol. 8, August 2026

The Write Kind of Memory: A Functionalist Critique of the Extended Mind Hypothesis

Tessa Popple[1]

University of Queensland

Abstract

This paper offers a functionalist critique of the Extended Mind Hypothesis as advanced by Andy Clark and David Chalmers in response to the question: "where is my mind?" While accepting the core functionalist premise that cognitive status is determined by causal role rather than location, it argues that the analogical Otto/Inga case fails to establish genuine functional parity. Thus, the Extended Mind Hypothesis fails. The argument draws on contemporary accounts of memory as a reconstructive and epistemically regulative process, supplemented by Eric Schwitzgebel's work on the unreliability of introspection, to argue that biological memory possesses functionally significant features absent from Otto's notebook. The paper concludes that despite surface similarities in the Otto/Inga analogy, a disanalogy can be drawn when the functionalism Clark and Chalmers depend on is truly taken seriously.

[1] *Tessa Popple is a third-year undergraduate student who is undertaking an extended major in philosophy at the University of Queensland. Her interests include the philosophy of mind and cognitive science, epistemology, and philosophy of logic.*

§ 1. From Extension to Objection

Contemporary philosophy of mind is increasingly shaped by the suspicion that the boundaries of the mental are far less clear than once assumed. Traditional images of an inner, self-contained mind have given way to models that emphasise embodiment, environmental coupling, and the pervasive role of external resources in cognition. Against this backdrop, Andy Clark and David Chalmers' (C&C) extended mind hypothesis (EMH) proposes that cognitive processes are not confined behind the skin-and-skull barrier but can extend into the world through our routine use of tools.[2] Their central case is the now-canonical analogy between Otto and Inga: Inga recalls information from biological memory, while Otto, who suffers from Alzheimer's disease, consults a notebook in which he has recorded the same information. When the notebook is reliably available, automatically trusted, and seamlessly integrated into Otto's activity, C&C argue that it plays the same functional role as memory, and should therefore be counted as part of his cognitive system.[3] The force of this claim depends on a commitment to functionalism: that what makes a process cognitive is the role it plays within a system, rather than its physical realisation or location.[4] If Otto's notebook performs the same function as Inga's memory, then denying it cognitive status would amount to privileging the inside of the head in a theoretically unmotivated way.[5] The EMH thus presents itself not as a radical departure from familiar assumptions, but as a natural consequence of taking functionalism seriously. In this paper, I argue that the Otto/Inga analogy fails to establish sufficient functional parity to support the EMH precisely because functionalism is taken seriously. The problem lies not in the appeal to external resources as such, but in the account of memory on which the analogy depends. C&C treat memory as if it were a passive store of information, whose role is exhausted by the reliable retrieval of stored content. Biological memory extends beyond storage, functioning as a reconstructive capacity that operates under conditions of uncertainty.[6] Its characteristic fallibility — a tendency to distort, misrepresent, and revise — is not an incidental defect, but a functionally essential feature of how it supports reasoning, learning, and action.[7] The appeal to subjectivity in this context is functional rather than phenomenological. It concerns the way memory regulates an agent's epistemic stance by presenting information with varying degrees of confidence, salience, and authority, not the qualitative character of remembering itself. Drawing on empirical work in the neuropsychology of memory and Eric Schwitzgebel's account of introspection, I argue that this reconstructive and epistemically regulative character is constitutive of biological

[2] Clark, Andy, and David Chalmers (1998), 'The Extended Mind', *Analysis* **58**: 7.
[3] Clark and Chalmers, *The Extended Mind*, 2–16.
[4] Clark and Chalmers, *The Extended Mind*, 8.
[5] Clark and Chalmers, *The Extended Mind*, 8.
[6] Schacter, Daniel L (2012) 'Adaptive Constructive Processes and the Future of Memory', The American Psychologist **67**: 603–13.
[7] Bartlett, Frederic C (1995) 'Perceiving, Recognising, Remembering', in *Remembering*: 186–96: Cambridge University Press.

remembering in a way that is not captured by Otto's notebook. I conclude that even if the mind extends, Otto and Inga do not provide sufficient evidence of it.

§ 2. Mind the Gap: Cognition Beyond the Skull

C&C's EMH is motivated by a substantive commitment to functionalism: the view that what makes a state mental is not what it is made of, nor where it is located, but the role it plays within a system of inputs, internal processes, and behavioural outputs. On this approach, mental states are individuated by their causal roles — how they are formed, how they interact with other states, and how they guide action — rather than by their physical realisation. While functionalism can encompass a range of positions, C&C rely on the broadly shared commitment that cognitive states are individuated by their functional roles rather than by the material in which they are realised. This allows for the possibility that the same mental state could, in principle, be realised in different forms, provided the relevant functional organisation is preserved. The EMH extends this commitment by rejecting the assumption that such functional organisation must be realised entirely within the biological boundaries of the individual. This motivates what is known as the parity principle for C&C: if a process outside the head plays the same explanatory role as one inside it, then to deny it cognitive status on the basis of location is arbitrary and theoretically unmotivated.[8] The principle is designed to block what they take to be an unjustified 'skin-and-skull bias' in our intuitions about the mind.[9] If we are willing to count a process as cognitive when it is realised neurally, consistency demands that we extend the same status to non-neural processes that perform the same function. This position reflects a broader reorientation within cognitive science, in which agents are understood not as self-contained information processors but as systems dynamically coupled to their environments. Cognition, on this view, is not exhausted by what happens internally, but is distributed across brain, body, and world. Human problem-solving routinely involves the active manipulation of external resources such as using a calculator to perform complex arithmetic, rearranging symbols on paper to solve a problem, or relying on digital devices to store and organise information. These practices are not merely convenient aids but are deeply integrated into our cognitive routines, shaping how tasks are structured and how solutions are reached. The EMH thus promises a convincing continuity between inner processes and our pervasive use of tools, presenting cognitive extension as a natural consequence of taking functionalism and our everyday practices seriously.

Traditionally, philosophy of mind has regarded cognitive processes as occurring wholly within the individual, whether understood in terms of brain states,

---

[8] Clark and Chalmers, *The Extended Mind*, 8–9.
[9] Clark and Chalmers, *The Extended Mind*, 14.

representations, or subjective experience.[10] This is the internal view of cognition which externalism rejects, arguing that the contents of our thoughts can depend in part on factors outside the individual. One such challenge to internalism is found in Hilary Putnam's *The meaning of 'meaning'*.[11] Through his Twin Earth thought experiment, Putnam argues that the meanings of our terms are not determined solely by our internal psychological states, since two individuals could be psychologically identical while nevertheless referring to different substances when they use the word "water".[12] On this view, features of the external environment partly determine mental content, leading Putnam to conclude that "meanings just ain't in the head."[13]

Whereas Putnam emphasises dependence on the physical environment, Tyler Burge develops this in *Individualism and the Mental* by arguing that what an individual thinks is partly determined by the linguistic community in which they participate.[14] He presents a thought experiment that aims to show that two individuals could be physically and psychologically identical, yet differ in what they think due solely to differences in how terms are used in their respective linguistic communities.[15] On this view, referred to as *passive* externalism, the environment plays a constitutive role in fixing the contents of mental states, but does so without locating those states themselves: beliefs, thoughts, and intentions remain states of the subject, even if what they are about depends on external factors.[16]

C&C both extend and depart from this framework by hypothesising *active* externalism. This goes further than the argument that the environment helps determine what we think, instead asserting that it can be directly involved in the processes by which we think.[17] Whereas passive externalism leaves cognition located within the individual while allowing its content to be environmentally shaped, C&C challenge the assumption that cognitive processes must be confined within the boundaries of skin and skull: the claim is not that the world helps determine what we think, but that it can literally participate in some of the thinking.[18] The EMH therefore asks us to accept not just that cognition is environmentally dependent, but that the environment itself forms part of cognition.

[10] See: Descartes, René (1998) *Discourse on Method and Meditations on First Philosophy*, 4th edition, Donald Cress, trans. Hackett; Fodor, Jerry A (1975) *The Language of Thought. Language and Thought Series*. Crowell, Adams, Fred, and Kenneth Aizawa (2001) 'Why the Mind Is Still in the Head', in Philip Robbins and Murat Aydede, eds., T*he Cambridge Handbook of Situated Cognition*: 78–95. Cambridge University Press.
[11] Putnam, Hilary (2011) 'The Meaning of 'Meaning'', in Hilary Putnam *Mind, Language and Reality*: 215–71. Cambridge University Press.
[12] Putnam, *The meaning of 'meaning'*, 223–26.
[13] Putnam, *The meaning of 'meaning'*, 227.
[14] Burge, Tyler (1979) 'Individualism and the Mental', *Midwest Studies in Philosophy* **4**: 73–121.
[15] Burge, *Individualism and the Mental*, 94.
[16] Burge, *Individualism and the Mental*, 79.
[17] Clark and Chalmers, *The Extended Mind*, 9.
[18] Clark and Chalmers, *The Extended Mind*, 9.

§ 3. An Analogy from Memory to Memo

This is the backdrop against which C&C introduce their central thought experiment, designed to make the extended mind hypothesis intuitive and sensitive to our daily practices. Having argued that cognition often involves the active use of environmental resources — such as manipulating objects, writing, or rearranging external symbols — they turn from general cases of problem-solving to more specific cases of belief and memory. Their aim is to show that external resources can play the same functional role as internal cognitive states and therefore qualify as constituents of cognition. Where sufficient functional parity exists, differences in physical location do not justify treating one process as cognitive and the other as merely supportive. To this end, they contrast two cases analogically.[19] Inga is an ordinary agent with standard cognitive capacities. Upon hearing that an exhibition is on at the Museum of Modern Art, she forms the intention to visit and recalls from her biological memory that the museum is on 53rd Street. Prior to recall, the information exists as a standing belief stored in memory, available when needed to guide action. When she retrieves it, she acts on it without hesitation and walks to the appropriate location. Otto, by contrast, suffers from Alzheimer's disease and cannot reliably store or retrieve information biologically. To compensate, he relies on an external artifact: a notebook in which he records information as he encounters it. When Otto learns about the exhibition, he writes down the address. Later, when he decides to visit, he consults the notebook, finds the relevant entry, and proceeds accordingly. The notebook is used by Otto as a constant and trusted resource — "the information is reliably there when needed, available to consciousness and available to guide action".[20] C&C argue that once these conditions are in place, there is no principled difference between the two cases. In both cases, the information is reliably available, easily accessible, and directly guides action in response to the agent's goals. To deny that Otto believes the museum is on 53rd Street, while granting that Inga does, would be to privilege the location of the information as inside the head over its functional role. This is precisely what their parity principle is designed to reject.

To avoid 'cognitive bloat', a concern expressed by critics of the EMH in which the boundaries of the mind become too permissive, C&C identify a set of conditions under which an external resource qualifies as part of a cognitive process: it must be reliably available, easily accessible, automatically endorsed, and integrated into the agent's ongoing activity.[21] [22] Otto's notebook, on this view, plays the same functional role for him as biological memory plays for Inga. The analogy is therefore not between a person and a person-with-a-tool, but between two cognitive systems that differ only in the location of their informational resources. If functionalism is accepted, C&C conclude, then parity is achieved. The notebook therefore constitutes Otto's memory itself, extending cognition beyond the skin-and-skull boundary.

---

[19] Clark and Chalmers, *The Extended Mind*, 13–14.
[20] Clark and Chalmers, *The Extended Mind*, 13–14.
[21] Clark and Chalmers, *The Extended Mind*, 11.
[22] Sprevak, Mark (2019) 'Extended Cognition', *The Routledge Encyclopedia of Philosophy Online*: 1–15.

C&C anticipate a range of objections to the claim of parity in the analogy, many of which arise from the intuitive sense that Inga's case involves a more direct or genuine form of cognition than Otto's.[23] These objections take several forms. One concerns access: Inga appears to enjoy a privileged, immediate relation to her memory, whereas Otto must consult an external object. Another concerns mode of access: Otto's interaction with the notebook is perceptual as he reads what is written, whereas Inga's access to memory is often characterised as introspective or non-perceptual. A further objection targets the nature of belief itself, suggesting that only occurrent mental states — those currently present to consciousness — qualify as genuine beliefs. In this case, Otto's reliance on a stored notebook entry would count only as dispositional or derivative. Finally, there are concerns about reliability and stability, since Otto's notebook might be lost, damaged, or unavailable in ways that biological memory typically is not.[24] Despite granting the intuition as to which these objections arise, C&C argue that they ultimately fail to identify a principled difference between the cases. Differences in access — perceptual versus introspective — are dismissed as variations in the mode of information retrieval rather than differences in its functional role. Likewise, concerns about reliability are treated as matters of degree rather than kind: biological memory is itself fallible, subject to interference, decay, and disruption, and thus cannot claim a privileged status on this basis alone. Even the appeal to occurrent belief is rejected as too restrictive, since much of what we ordinarily count as belief consists in standing states that are not continuously present to consciousness but are nonetheless available to guide action when required. In each case, C&C's strategy is to show that the apparent asymmetries between Otto and Inga are superficial — differences in medium, phenomenology, or implementation — rather than differences in the underlying cognitive role played by the relevant information.

Formally, the argument is both conditional and analogical. It proceeds from a functionalist premise: namely, that cognitive status is determined by the role a process plays within a system, and that it applies to an analogy in which an internal and an external process appear to realise that role equally well. If functionalism is true, and if Otto's notebook is functionally equivalent to Inga's biological memory, then the information stored in the notebook must count as a belief or memory state. One influential line of resistance, developed by Fred Adams and Kenneth Aizawa, rejects the sufficiency of functional equivalence altogether.[25] They argue that C&C's account lacks a principled criterion for distinguishing genuinely cognitive processes from merely causal accompaniments and propose instead that cognition is marked by features such as the presence of non-derived representations and specific kinds of neural processing.[26] On this view, external artifacts like notebooks may play an important role in facilitating cognition but do not themselves constitute cognitive processes, since their representational content is derivative on human interpretation.

[23] Clark and Chalmers, *The Extended Mind*, 14–16.
[24] Clark and Chalmers, *The Extended Mind*, 14–16.
[25] Adams and Aizawa, *Why the Mind Is Still in the Head*, 78–79.
[26] Adams and Aizawa, *Why the Mind Is Still in the Head*, 80.

This objection thus targets the functionalist premise directly, seeking to reinstate a boundary between the cognitive and the non-cognitive that is not captured by parity of role alone.

§ 4. Reconstructing Memory, Deconstructing Parity

We need not reject functionalism in order to resist C&C's conclusion. Indeed, C&C themselves acknowledge their argument leaves open a different strategy of objection: namely, to accept functionalism but deny relevant functional equivalence has been established. As they explicitly concede, to "provide substantial resistance, an opponent has to show that Otto's and Inga's cases differ in some important and relevant aspect."[1][27] This shifts the burden of the debate. The question is no longer whether cognition must be internal, nor whether functionalism is correct, but whether the Otto/Inga analogy genuinely satisfies the conditions of parity. Parity stands only if no deeper functional disanalogy can be shown.

Otto's and Inga's cases do differ in an important and relevant way once a more nuanced account of memory is adopted. C&C's argument succeeds only if memory is understood as a passive store of information whose function is exhausted by the reliable retrieval of stored content. Biological memory is instead a perspective-bound, reconstructive capacity that retrieves information under conditions of uncertainty, salience, and confidence.[28] Memory is fallible *as memory*, distorting and misrepresenting information while often presenting itself as authoritative.[29] As Loftus and Pickrell observe, "the memory error occurs because details of experienced events or imagined events are integrated with inferences and other elaborations," a consequence of the ordinary mechanisms through which memories are formed and retrieved.[30] These failures are not occasional defects, but internal to the way remembering functions for a subject who stands in a first-person relation to their past. Remembering functions through the continual updating, reorganisation, and evaluation of representations in light of present circumstances. This functional organisation enables flexible reasoning and adaptation while also generating the possibility of distortion. Error therefore arises from the ordinary operation of remembering itself. Inga can be mistaken in recalling an address, certain when she should doubt, or unsure when she is in fact right. This is central to what memory does. Otto's notebook, by contrast, is static and epistemically transparent. It does not reinterpret its contents, reorder them in response to contextual cues, or present them with varying degrees of confidence. It is just as "sure" of what Otto ate for lunch five years ago as it is of the museum's address — a mistaken notebook entry is simply an incorrect record.

---

[27] Clark and Chalmers, *The Extended Mind*, 14.
[28] Schacter, *Adaptive Constructive Processes and the Future of Memory*.
[29]Loftus, Elizabeth F, and Jacqueline E Pickrell (1995) 'The Formation of False Memories', *Psychiatric Annals* **25**: 720–25.
[30] Loftus and Pickrell, *The Formation of False Memories*, 724.

Consider a conversation in which Inga's friend reminds her of an event from years ago. Her recollection may subsequently change as the reconstructive mechanisms of memory integrate new contextual information into the remembered event. The resulting revision is produced by memory's ordinary operation rather than by damage or malfunction. By contrast, if Otto's notebook now records that the guest arrived late, this is because the entry has been altered or because Otto originally recorded it differently. In this respect, Otto may even possess a more stable informational resource than Inga, since the notebook preserves fixed content without the reconstructive distortions characteristic of memory.

Biological memory is a norm-governed, perspectival capacity whose characteristic fallibility is internal to its function. Its role includes regulating an agent's epistemic stance by selecting, weighting, and reconstructing representations in light of context, alongside the storage and retrieval of information. Those same functional processes inevitably introduce the possibility of distortion, making fallibility constitutive of memory rather than an accidental defect. The notebook lacks this core feature. This objection does not rest on the claim that Otto's notebook lacks reliability, availability, or integration, nor that it differs merely in medium or location. The relevant distinction concerns the functional organisation of remembering rather than the phenomenology of memory; the appeal to subjectivity here is functional. Biological memory presents information, while also regulating epistemic stance by producing representations graded in confidence, shaped by context, and open to revision. Retrieval is therefore inseparable from the evaluation and reconstruction of remembered content. The same processes that make a memory more or less salient, tentative, or authoritative also integrate new information and can alter the remembered representation itself. By contrast, consulting a notebook retrieves a comparatively stable record whose epistemic significance is determined through subsequent interpretation. Any change to the stored content requires a distinct act of revision rather than emerging through the ordinary process of retrieval itself. Otto and Inga therefore differ in an important and relevant way.

Eric Schwitzgebel's work on the unreliability of introspection provides further support for this claim by challenging a deeply entrenched assumption in the philosophy of mind: that we enjoy a privileged and largely authoritative access to our own mental states. Traditionally, introspection has been treated as a kind of epistemic bedrock, a domain in which error is minimal and self-knowledge secure.[31] Schwitzgebel objects to this assumption, arguing that even under favourable conditions of careful reflection, introspection is systematically and pervasively unreliable. He contends that we are often "ignorant and prone to error" about our own ongoing conscious experience, not only in marginal or difficult cases, but even with respect to its most basic and pervasive features.[32] For example, he argues that most people initially judge themselves to possess a broad field of stable, richly

[31] See Descartes, *Discourse on method ; and, Meditations on first philosophy.*; Shoemaker, Sydney (1963) *Self-Knowledge and Self-Identity* Cornell University Press.
[32] Schwitzgebel, Eric (2008) 'The Unreliability of Naive Introspection', *The Philosophical Review* **117**: 245–73.

detailed visual experience. However, when they undertake simple introspective exercises—such as maintaining fixation while attending to peripheral vision—they frequently revise this judgment, recognising that the apparent richness of visual experience was substantially overstated.[33] For Schwitzgebel, this demonstrates that even careful reflection can expose systematic errors in our understanding of our own conscious experience. This unreliability is not limited to judgments about the causes of our mental states, or about dispositional attitudes such as character traits or motivations, where error might be expected. Schwitzgebel's examples are deliberately mundane and wide-ranging. We struggle to characterise the nature of our emotional experience — whether emotions have a stable phenomenological core, how they are spatially or bodily located, or even whether we are undergoing them at all. Similarly, we misjudge the structure of our visual experience, overestimating the clarity and detail of the visual field and failing to appreciate the limited and shifting nature of perceptual attention.[34] These mistakes are common and persistent features of ordinary cognition. Even when we reflect sincerely and attentively, our judgments about our own experience are unstable, revisable, and often incorrect. Schwitzgebel concludes that mental states are not self-presenting in the way many philosophical accounts assume.[35] Our access to them is mediated, interpretive, and subject to the same kinds of error, uncertainty, and revision that characterise our engagement with the external world. To have a belief or a memory is thus not to stand in a relation of transparent self-knowledge, but to occupy a position that admits of misjudgement about one's own epistemic state. This supports the present argument by reinforcing the idea that fallibility is not a structural feature of cognition: the same processes that enable access to mental states also necessitate the possibility of error in accessing them. The question, then, is whether features of this picture of memory can be attributed to Otto's notebook.

As such, C&C may respond that this line of argument fails to establish a principled disanalogy. That biological memory exhibits a distinctive form of fallibility does not, on their view, mark a relevant difference, since Otto's notebook is not perfectly reliable either: it can be lost, damaged, misread, or even tampered with. In this respect, both systems are susceptible to error, and the source of that error — whether internal or external — does not by itself determine cognitive status. What matters, they argue, is not perfect reliability, nor the presence of a distinctive phenomenology, but whether a resource is sufficiently available, accessible, and integrated into the agent's cognitive routines to play the appropriate action-guiding role.[36] On this view, the relevant standard is one of functional integration rather than intrinsic character. Biological memory is itself imperfect, subject to interference, decay, and distortion, yet it nonetheless counts as memory because it is reliably available when needed and

[33] Schwitzgebel, *The Unreliability of Naive Introspection*, 255–56.
[34] Schwitzgebel, *The Unreliability of Naive Introspection*, 265–67.
[35] Schwitzgebel, *The Unreliability of Naive Introspection*, 268.
[36] Clark, Andy (2008) *Supersizing the Mind : Embodiment, Action, and Cognitive Extension. Philosophy of Mind*. Oxford University Press.

is seamlessly incorporated into ongoing cognition. If Otto's notebook is similarly dependable, then it satisfies the same functional criteria. That its failures arise through external damages rather than through reconstructive distortion is treated as a difference in implementation, rather than a difference in role. From this perspective, differences in implementation or location do not by themselves justify denying parity, since doing so would risk reinstating the very skin-and-skull boundary that the EMH challenges.

This reply mislocates the source of the distinction. The contrast is not between systems that can err and systems that cannot, but between systems whose *normal operation* includes epistemic risk and those whose errors arise only from malfunction. Fallibility due to loss, corruption, or damage is merely contingent: it reflects what can happen to any physical object. Fallibility due to reconstruction, by contrast, is constitutive of memory. Remembering reconstructs representations of past events through context-sensitive processes that operate from the subject's epistemic perspective. In memory, error is produced by the same mechanisms that make recall possible at all: in a system involving a notebook, error arises through failures in interpretation, retrieval, or integration, rather than from the operation of the storage medium itself.[37] This point survives the reply that error may occur at the level of the coupled system. Otto may misread, misinterpret, or misapply what is written just as he may misjudge his own memories. But in such cases, the source of the error lies in the cognitive processing brought to bear on the stored information, not in the way the information is stored or maintained. The notebook contributes fixed content; it does not itself generate graded, revisable, or confidence weighted representations. This point can be maintained even if we grant that the relevant cognitive system is not Otto alone but the coupled system of Otto-plus-notebook (O+N). The issue is not whether interpretation occurs within the system — both biological memory and O+N systems involve processes of interpretation, reconstruction, and error — but where and how those processes are realised. In biological memory, reconstruction is internal to the mechanisms of storage and retrieval themselves: the processes that make recall possible such as selection, abstraction, and contextual modulation, also generate distortion. Error is thus constitutively produced within memory's ordinary operation.

In the O+N system, the notebook contributes stable propositional content, and all reconstruction occurs only at the level of the subject's interaction with that content. The system may misinterpret, misapply, or misremember what is written, but these errors arise from downstream processing applied to a stable informational resource. The result is a division of labour absent in biological memory: storage remains static, while epistemic regulation is imposed upon it. For this reason, even at the level of the coupled system, O+N does not realise a form of memory whose normal operation inherently generates and regulates epistemic risk. A system that retrieves fixed content and then interprets it is not performing the same epistemic labour as

[37] Lentoor, Antonio G (2023) 'Cognitive and Neural Mechanisms Underlying False Memories: Misinformation, Distortion or Erroneous Configuration?', AIMS neuroscience **10**: 255–68.

one in which retrieval itself is reconstructive, graded, and revisable. The notebook may participate in a cognitive system that errs, but it does not contribute to a system whose errors are generated by the very processes that make remembering possible.

Recent empirical evidence from the field of neuroscience provides further support for this picture of fallible reconstruction and its functional role. Work on fuzzy trace theory, for example, suggests that memory does not store a single, unified representation of past events, but operates through the parallel encoding of two distinct kinds of traces: verbatim traces, which preserve surface-level details, and gist traces, which encode the general meaning or structure of an experience.[38] The reconstructive character of remembering depends, in part, on the availability of multiple representational formats. Reyna and Brainerd observe that "subjects encoded more exact representations that... were not used much in reasoning... Thus, a range of representations varying in precision, fuzzy-to-verbatim continua, were available for reasoning."[39] While verbatim traces support precise and accurate recall, they are relatively fragile and decay quickly. Gist traces, by contrast, are more durable and play a central role in reasoning, generalisation, and decision-making.[40] Crucially, however, it is these very gist-based processes that give rise to systematic distortions and false memories. This is not an accidental by-product of an otherwise reliable system, but a consequence of how memory typically functions. In contexts such as survival processing, where information is encoded in a deeper and more meaning-laden way, gist traces are strengthened, leading not only to improved recall but also to an increased likelihood of confidently misremembering related but non-presented information.[41] The same mechanisms that enable abstraction—allowing us to extract patterns, form categories, and apply past experience to new situations — also generate error. Memory's role in learning and adaptation therefore depends on its inherent tendency to misremember; without the ability to move beyond exact detail and represent the general significance of events, cognition would be far less flexible and far less useful.[42] The challenge for the EMH remains not whether Otto can reliably access information, but whether the notebook participates in the kinds of reconstructive processes that appear constitutive of ordinary memory.

What emerges from this picture of memory is not a passive store of information, but an epistemically regulative system—one that presents remembered information with varying degrees of confidence, salience, and authority. These features are functionally significant as they guide how information is used in reasoning, when it is revised, and how it is integrated with other beliefs. A memory may be tentative or vivid, marginal or central, easily overridden or strongly action-guiding, and these

---

[38] Bialer, Daniel M., Minyu Chang, C J Brainerd, and Valerie F Reyna (2024) 'A Fuzzy Trace Theory Account of Survival Processing', in *Interdisciplinary Perspectives and Advances in Understanding Adaptive Memory*: 402–24. Oxford University Press.
[39] Reyna, V F, and C J Brainerd (1995) 'Fuzzy-Trace Theory: An Interim Synthesis', *Learning and Individual Differences* **7**: 1–75.
[40] Bialer et al., *A Fuzzy Trace Theory Account of Survival Processing*, 406.
[41] Bialer et al., *A Fuzzy Trace Theory Account of Survival Processing*, 418.
[42] Bialer et al., *A Fuzzy Trace Theory Account of Survival Processing*, 422.

modulations are part of what it is to remember at all. The appeal to subjectivity, in this context, is therefore not phenomenological but functional: it concerns the way information is processed and deployed within cognition, rather than how it feels. By contrast, Otto's notebook lacks any such internal structure. It preserves information in a fixed, non-modulating form, and whatever variation in confidence, relevance, or interpretation occurs does so only at the level of Otto's subsequent engagement with it. The notebook does not itself weigh, reinterpret, or reconstruct its contents; it does not generate error through its ordinary operation, but only insofar as it is misread, damaged, or misapplied. For this reason, although the notebook may contribute to a cognitive system that is fallible, it does not realise a form of memory whose characteristic fallibility is constitutive of its function. Otto's notebook can fail as a tool, but it cannot fail in the way a mind does.

§ 5. No Charity from Parity

To see what parity would actually require, we can ask what Otto's notebook would have to be like if it genuinely played the same functional role as biological memory. Reliable storage and accessibility alone would be insufficient. It would need to participate in the same reconstructive, epistemically regulative processes that generate both recall and distortion in ordinary remembering. The entry "MoMA is on 53rd Street" could not simply remain a stable record until deliberately revised. Rather, its content would need to become more or less authoritative, salient, or action-guiding through the ordinary process of retrieval itself, as Otto's circumstances changed. Likewise, revision would need to arise through the same mechanisms that make remembering possible, rather than through the separate act of crossing out an entry, annotating it, or replacing it with a new one. The notebook would therefore need to generate reconstruction as part of its ordinary operation, integrating new contexts into the remembered representation itself rather than preserving fixed content until externally modified.

Otto can, of course, update the notebook when presented with conflicting evidence. However, this requires a distinct act of inscription that produces a new record rather than a transformation of the remembered representation through retrieval itself. The notebook preserves what has been written until it is deliberately altered, and retrieval simply presents that same stored content for subsequent interpretation. Relevant parity therefore requires more than an external record whose contents can be rewritten. It requires a memory system in which retrieval is itself reconstructive: the processes that make remembering possible also integrate contextual information, regulate confidence, and revise the remembered representation. In biological memory, storage, retrieval, and reconstruction are constitutive aspects of a single functional process. An ordinary notebook, by contrast, stores stable propositional content that may subsequently be interpreted, annotated, or overwritten, but those operations are functionally distinct from the retrieval of the stored information itself.

The notebook therefore participates in remembering without realising the same reconstructive organisation that characterises biological memory.

The extended mind hypothesis is compelling in its observation that we think with the world. Tools, symbols, and environments are not peripheral to cognition but are deeply entangled with it, shaping how we reason, remember, and act. C&C are right to resist an unreflective attachment to the boundaries of the skull and to emphasise the continuity between internal processes and our everyday reliance on external resources. However, the move to cognitive status is too quick. The argument from parity depends not simply on the presence of reliable, integrated external resources, but on their realising the same kind of cognitive process as their internal counterparts. It is here that the analogy between Otto and Inga fails. Biological remembering is not a matter of passively retrieving stored information, but of reconstructing the past under conditions of uncertainty, guided by processes that regulate confidence, salience, and revision. Its characteristic fallibility is not a defect to be engineered away, but a feature internal to its function. Otto's notebook cannot participate in this kind of process. However indispensable it may be to his cognitive life, it remains a system that stores and delivers fixed content, leaving all epistemic regulation to the agent who consults it. Its failures arise through breakdown, misinterpretation, or misuse, rather than from the ordinary operation of a reconstructive system. As such, it supports cognition without itself instantiating a memory state. The relevant standard is not whether a resource is used in thinking, but whether it realises the same kind of epistemic work. In the case of Otto and Inga, this standard is not met. If the mind extends, Otto's notebook is not where it ends up.

References

Adams, Fred, and Kenneth Aizawa (2001) 'Why the Mind Is Still in the Head', in Philip Robbins and Murat Aydede, eds., T*he Cambridge Handbook of Situated Cognition*: 78–95. Cambridge University Press

Bartlett, Frederic C (1995) 'Perceiving, Recognising, Remembering', in *Remembering*: 186–96: Cambridge University Press.

Bialer, Daniel M., Minyu Chang, C J Brainerd, and Valerie F Reyna (2024) 'A Fuzzy Trace Theory Account of Survival Processing', in *Interdisciplinary Perspectives and Advances in Understanding Adaptive Memory*: 402–24. Oxford University Press.

Burge, Tyler (1979) 'Individualism and the Mental', *Midwest Studies in Philosophy* **4**: 73–121. DOI: 10.1111/j.1475-4975.1979.tb00374.x

Clark, Andy (2008) *Supersizing the Mind : Embodiment, Action, and Cognitive Extension. Philosophy of Mind*. Oxford University Press.

Clark, Andy, and David Chalmers (1998) 'The Extended Mind', *Analysis* **58**: 7–19. DOI: 10.1111/1467-8284.00096

Descartes, René (1998) *Discourse on Method and Meditations on First Philosophy*, 4th edition, Donald Cress, trans. Hackett.

Fodor, Jerry A (1975) *The Language of Thought. Language and Thought Series*. Crowell.

Lentoor, Antonio G (2023) 'Cognitive and Neural Mechanisms Underlying False Memories: Misinformation, Distortion or Erroneous Configuration?', AIMS neuroscience **10**: 255–68. DOI: 10.3934/Neuroscience.2023020

Loftus, Elizabeth F, and Jacqueline E Pickrell (1995) 'The Formation of False Memories', *Psychiatric Annals* **25**: 720–25. DOI: 10.3928/0048-5713-19951201-07

Putnam, Hilary (2011) 'The Meaning of 'Meaning'', in Hilary Putnam *Mind, Language and Reality*: 215–71. Cambridge University Press.

Reyna, V F, and C J Brainerd (1995) 'Fuzzy-Trace Theory: An Interim Synthesis', *Learning and Individual Differences* **7**: 1–75. DOI: 10.1016/1041-6080(95)90031-4

Schacter, Daniel L (2012) 'Adaptive Constructive Processes and the Future of Memory', The American Psychologist **67**: 603–13. DOI: 10.1037/a0029869

Schwitzgebel, Eric (2008) 'The Unreliability of Naive Introspection', *The Philosophical Review* **117**: 245–73. DOI: 10.1215/00318108-2007-037

Shoemaker, Sydney (1963) *Self-Knowledge and Self-Identity* Cornell University Press.

Sprevak, Mark (2019) 'Extended Cognition', *The Routledge Encyclopedia of Philosophy Online*: 1–15. DOI: 10.4324/9780415249126-V049-1

Undergraduate Philosophy Journal of Australasia www.upja.online
Vol. 8, August 2026

Epistemically Contaminated Non-Binary Aesthetic Self-Expression: A New Form of Hermeneutical Injustice

Bowen Northrop[1]

Cardiff University

Abstract

This paper argues that aesthetic self-expression, the investigation and communication of one's interests, tastes, beliefs and passions, can be a useful method of gaining self-knowledge. Further, I argue that when this exercise is wrongfully inhibited, this can be an instance of hermeneutical injustice, the unjust obscuration of one's experiences from collective understanding, to follow Fricker's analysis. Specifically, my focus is on how cognitive hermeneutical injustice, the unjust inhibition of one's ability to *make sense* of their experiences, might arise from having one's aesthetic self-expression compromised. My core case centres around how non-binary people experience barriers to self-expression and how this in turn may wrongfully affect their ability to make sense of their experiences. I argue that this hindered self-expression is the product of epistemic contamination – the conceptual load of an aesthetic product or practice limiting a person's ability to engage freely in self-expression. This discussion concludes that Epistemically Contaminated Aesthetic Self-Expression (ECASE) is not merely an instance of hermeneutical injustice specific to non-binary people but is a form of hermeneutical injustice that poses a wider risk.

[1] *Bowen is a recent Philosophy BA graduate from Cardiff University. Their philosophical interests include aesthetic injustice, social epistemology and feminist philosophy.*

§ 0. Introduction

The focus of non-binary hermeneutic literature has been on the presence of traditional epistemic barriers towards non-binary people and its communicative harm[2]. Notably, this has included the hermeneutical injustices prompted by structural grammatical exclusion, gatekeeping via intelligibility in access to gender affirming healthcare, and category invalidation[3]. This current literature has done the groundwork in establishing an understanding of the dominant binary gender framework as a hostile epistemic environment. This is an environment that significantly hinders one's ability to know something, an environment that might lead to hermeneutical injustice. Following this, I want to highlight the *cognitive* hermeneutical injustice that can follow from the same poor epistemic conditions, and in doing so, pertaining to the experiences of non-binary people, bring into light a novel form of cognitive hermeneutical injustice.

Specifically, my aim in this paper is to introduce epistemic contamination of aesthetic self-expression *qua* aesthetic objects and aesthetic environments as a form of cognitive hermeneutical injustice. In section 1, I will explain what hermeneutical injustice is and the kinds relevant to this essay. In section 2, I will establish what I mean by aesthetic self-expression (ASE) and in which conditions this appears to have epistemic value. In section 3, I will argue for the main claims in 3.1, that non-binary people are in a position of epistemic hostility, and in 3.2, that the epistemic contamination that follows from this hostility negatively affects the epistemic potential of ASE. Finally, I conclude that Epistemically Contaminated Aesthetic Self-Expression (ECASE) is a form of cognitive hermeneutical injustice, which I discuss in section 4.

Firstly, however, I will explain some key terms that will be used throughout the paper. 'Epistemic' relates to knowledge and how we come to know things. This paper focuses on knowledge about oneself and one's identity. 'Hermeneutics' describes the study of interpretation. Here, it is used in reference to one's understanding and ability to interpret *oneself*, as inspired by Miranda Fricker's work on hermeneutical injustice[4]. The kind of 'self-expression' I discuss is directly tied up to ideas of the 'aesthetic'- the realm of art and related sensory experiences[5]. The aesthetic self-expression to which this paper refers is the aesthetic practices and

---

[2] 'Non-binary' in this paper refers to all gender identities that fall outside of the male/female binary standard. I thus use it as an umbrella term, rather than a singular gender category.

[3] Carmona, Carla, (2022). Binarism Grammatical Lacuna as an Ensemble of Diverse Epistemic Injustices, *Social Epistemology*, 37(3), pp.339–363; Porkkala, Siiri (2024) Hermeneutical Injustice, Nonbinary Gender Identities and Category Invalidation, *Social Epistemology*, 39(4), pp.1–11; Rosola, Martina, (2024) Linguistic Hermeneutical Injustice, *Social Epistemology*, 39(4), pp.1–11; Clanchy, Nick. (2024) *Tackling Hermeneutical Injustices in Gender-Affirming Healthcare*.

[4] Fricker, Miranda, (2007) Hermeneutical Injustice, In *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

[5] I keep my reference of the aesthetic intentionally vague and open here, to accommodate differing views on what makes something strictly aesthetic rather than affective. I merely use the phrase to gesture at that school of thought and bridge a connection between the aesthetic and the epistemic.

related products one engages with to communicate and investigate their like or affinity towards something. This can range from traditional aesthetics like making music to mundane self-expressive efforts, like making one's handwriting appear more feminine.

§ 1. Hermeneutical Injustice

This section outlines the explanatory groundwork necessary to establish the paper's main claim about *Epistemically Contaminated Aesthetic Self-Expression* (ECASE) as a form of hermeneutical injustice. Namely, we need to understand what hermeneutical injustice is and to specify the kind of hermeneutical injustice relevant to this discussion. As coined by Fricker, hermeneutical injustice describes 'the injustice of having some significant area of one's social experience obscured from collective understanding owing to hermeneutical marginalization.'[6]

As prompted by the distinction made by Goetze, we can unpack this definition by explaining hermeneutical injustice in terms of its primary harm: communicative or cognitive[7]. *Communicative* hermeneutical injustice is when 'the subject is rendered unable to make communicatively intelligible something which it is particularly in his or her interests to be able to render intelligible.'[8] To use Fricker's example, Carmita Wood was a victim of sexual harassment by her boss[9]. Subsequent trauma led to Wood developing chronic pain which forced her to quit her job. When filing for unemployment insurance, she did not have the term 'sexual harassment' available to her (both as a tick-box option on the unemployment forms nor in her vocabulary) to explain her reasons for having to quit. After describing her reasons for quitting as 'personal', her insurance claim was denied. The inability to communicate her experiences, separable from its harmful consequences relating to her employment status, is an instance of communicative hermeneutical injustice.

Cognitive hermeneutical injustice is when the subject cannot make sense of their own experiences, owing to a lack of appropriate conceptual resources[10]. These resources can include the words necessary to describe one's experience, or the conceptual frameworks to draw upon when making sense of something. Here, the unintelligibility of the experience is a problem for the subject themselves and, following Fricker's discussion, appears to relate to a harm to one's cognitive faculties, one's *ability* to understand a specific subject[11]. In Wood's case, this is a subsequent inability to make sense of the traumatic experience. This inability to place what happened to her into a wider concept of what sexual harassment is and

[6] Fricker, *Hermeneutical Injustice*, 158.
[7] Goetze, Trystan, S, (2018) *Hermeneutical Dissent and the Species of Hermeneutical Injustice*, *Hypatia*, *33*(1), 73-90.
[8] Fricker, *Hermeneutical Injustice*, 162.
[9] Fricker, *Hermeneutical Injustice*, 149-151.
[10] Goetze, *Hermeneutical Dissent and the Species of Hermeneutical Injustice;* Fricker, *Hermeneutical Injustice.*
[11] Fricker, *Hermeneutical Injustice.*

how she is not to blame may have caused severe psychic and emotional harm. This paper strictly discusses ECASE's cognitive harms, the harms of it being harder to make one's gender identity intelligible to oneself[12].

Both kinds of hermeneutical injustice require *hermeneutical marginalisation* as a prerequisite. A group is hermeneutically marginalised when they are in a position of unequal hermeneutic participation in regards to 'some significant area(s) of social experience.'[13] When a group is not being given ample power in developing shared cognitive resources within the social imaginary, particularly those that are significant to their social experiences, they are victims of hermeneutical marginalisation. In Wood's case, this is women's unequal participation in defining how the (mis)treatment of women is conceptualised in the workplace.

The product of this marginalisation is a lacuna in the collective understanding where ample resources to make sense of their experiences should lie. Each lacuna includes, in its commitment to the exclusion of those marginalised experiences, alternate explanations for those experiences. This appears one major way in which the 'whole engine of collective social meaning is effectively geared to keeping these obscured experiences out of sight.'[14] In Wood's case, the sexual harassment she experienced was considered harmless workplace flirting[15]. In the moment of realisation where Wood helped conceive the term 'sexual harassment', she didn't merely fill a missing hermeneutical lacuna but also pulled away a narrative blanket intended to conceal the true nature of that experience.

Finally, all kinds of hermeneutical injustice require that the epistemic harm done to that person is genuinely wrongful, rather than be the product of epistemic bad luck[16]. While it might be upsetting to not know what will happen when one dies or cognitively challenging to forget things as one gets older, these are only instances of epistemic bad luck. The lack of conceptual resources available to Wood were not epistemic bad luck, but the product of unjust hermeneutical marginalisation which resulted in a hermeneutical injustice.

The shared inherent harm across instances of hermeneutical injustice is damage to one's ability to make sense of their experiences and themselves. Often unknown and unnamed is the self-blame, confusion, and shame that follows from being unable to explain and understand what has happened to you and who you are. It's effect burrows deep into one's position not only as a social agent (in its ability to reinforce unjust power structures like the patriarchal dynamic existing in Wood's case by keeping her silent and at a hermeneutic disadvantage) but also in one's position as an individual, one's ability to see themselves clearly and be understood within a

[12] I suspect that communicative hermeneutical injustice also occurs in cases of ECASE, however, I am limited to the available discussion space here and must leave this to be discussed elsewhere.
[13] Fricker, *Hermeneutical Injustice*, 153.
[14] Fricker, *Hermeneutical Injustice*, 153.
[15] Fricker, *Hermeneutical Injustice*, 148-152; Susan, Brownmiller, (1990) *In Our Time: Memoir of a Revolution*, New York: Dial Press, 182.
[16] Fricker, *Hermeneutical Injustice*, 151.

social space. Being barred access to oneself in this way is part of what makes hermeneutical injustice so unjust.

Having explained the stakes and conditions of hermeneutical injustice, I can now begin to explain where aesthetic self-expression plays a role.

§ 1. The Epistemic Relevance of Aesthetic Self-Expression

This section seeks to establish this paper's first core claim: (playful) aesthetic self-expression functions as a key method for furthering self-understanding. I thus argue that aesthetic self-expression (ASE) is epistemically valuable, particularly while attempting to understand oneself when typical epistemic resources are scarce or otherwise unavailable.

'Self-expression' is often used in reference to the management and manipulation of our personal aesthetic presentations. For the purposes of this essay, 'self-expression' will relate to the following:

> Expressing one's genuine interest in or alignment with something; a way of signalling and engaging with something one likes/ stands for / is interested in / is a relevant feature of *who one is*, often through aesthetic means.

For instance, someone might engage in self-expression by adhering to the aesthetic norms of the music subculture they are a part of, such as a punk enjoyer wearing leather jackets with political patches. This communicates to others their identity and allows the person to engage with a culture/idea/norm that is important to them through aesthetic means. As such, we expect self-expression to be authentic and thus autonomous. From this, we see how one's understanding of the self can be reflected in one's aesthetic engagements and commitments.

However, this paper suggests that in many cases, such as the examples below, our ASE does not merely *reflect* who we are, but rather, is a key method for us to *develop our understanding* of who we are and/or make sense of our experiences:

> Case 1: Katie uses art as a means of therapy: 'Art is a tool that helps me manage some of the challenges created by life's changing circumstances. Making things is such a beautiful gift, even if it cannot entirely solve every issue, it can soothe and *help me process things* […] Get lost in some intentional *play*, and find some *clarity* and calm along the way.'[17]

> Case 2: A is a trans woman and member of the hijra community. In childhood, A felt conflicted about her gender identity, having had no one she felt she could confide in ('I could not talk to anyone about my confusions') and no way to explore her identity[18]. However, she recounts multiple instances where

[17] Chandler, Katie, (2019) "*Art Has Been a Healing Tool.*" Emphasis added.

[18] Revathi, A and V Geetha, (2010) *The Truth about Me: A Hijra Life Story*, London: Penguin Books. See 6.

> a more flexible aesthetic context allowed for revelatory aesthetic experiences. For example, she found a rare opportunity to play the part of a kurathi in the local Mariamman festival: 'To the world, it appeared that I was dressing up and playing a woman, but inside, I felt I was a woman. […] I looked at myself in the mirror several times, *astonished at what I had become*.'[19]. A talks about moments like these as crucial in cueing her into being a woman: 'I looked at myself in the mirror and felt a glow of pride. I did look like a woman. It was at that moment that I was convinced I was indeed one.'[20]

What these cases share is that some barrier to understanding was overcome in part due to an investigative and playful kind of aesthetic self-expression (ASE). In case 1, this barrier appears to be an internal hostile epistemic environment affected by her mental health issues. While Katie may have also benefited from traditional talking therapy, she says that the medium of art helps calm her into a position where she is better equipped to explore and understand herself. Seemingly for Katie, there is epistemic and aesthetic value in both the process and the cognitive results of art making.

In case 2, this barrier appears to be the result of an external hostile epistemic environment. This is shown in the lack of traditional epistemic resources available to her, such as listening to another trans person's testimony or talking her feelings through with someone directly. Similarly to Katie, A seems to have been able to use aesthetic expression as an epistemic tool to better make sense of her experiences, despite them being in different epistemic positions.

To engage in self-expressive aesthetic endeavours like these seems to require a level of emotional and physical safety, especially for those whose identities are (hermeneutically) marginalised. Even when someone is just trying out a new hairstyle, anxieties about ridicule, shame, embarrassment, concerns about cost, and such, may limit that person's ability to adopt new self-conceptions. These self-expressive exercises are, by nature, a little intimidating, as one is offering themselves up to potentially different and perhaps harmful treatment by others. This is particularly true of marginalised people, as their aesthetic self-expression is more likely to be scrutinised and stereotyped by those around them[21]. As such, ASE seems to require that one is assured of the safety and freedoms of their self-expressive exercises.

Further, both cases seem to involve a gradual revelatory experience, something investigative, cautious and playful, rather than a singular moment of realisation. Taking the former as the process of ASE's epistemic benefits seems not only more accurate to these cases, but in A's case, also avoids a pervasive stereotype, that trans people have this specific moment in childhood, usually a matter of trying on their mother's clothes in secret, which revealed to them a fully formed gender identity

[19] Revathi, *The Truth about Me: A Hijra Life Story,* 9-12. Emphasis added.
[20] Revathi, *The Truth about Me: A Hijra Life Story,* 16.
[21] Thank you to the UPJA team for this added point.

that had been lurking under the surface since birth[22]. Rather than frame A's experiences within this conception, I am discussing it as it appears to actually stand: one of many accruing moments in which the visual cues of her aesthetic self-expression and her reactions to it taught her about who she is: 'I experienced a *growing sense* of irrepressible femaleness, which haunted me, day in and day out.'[23]

These self-expressive revelations are important in navigating our identities and how we exist in relation to the rest of the social world. It makes one able to communicate to others who they are, understand how they relate to others with contrasting and shared identity markers, and develop their sense of personal identity. For example, wearing a suit might appear mundane, but really may help that person find a feeling of unity in a workplace, communicate to others the kind of work they do, and signal to themselves that this is a new chapter in who they are becoming. For those marginalised individuals already pushed out of the dominant social imaginary, finding and understanding one's place in the world seems particularly important for their epistemic and general wellbeing, and self-expression seems a means of doing so.

In understanding ASE's epistemic benefits as being gradually and intermittently released, rather than some grand and singular revelation, ASE is better epistemically accounted for and better suited for the topic at hand: self-development and realisation. This also aligns with Katie's testimony in Case 1: 'I like to think of my personal practice as an *ongoing relationship with my curiosity*, allowing me to both explore and to express.'; 'For me, there hasn't been one, defining "aha!" moment of appreciation, but rather small moments I've connected over time.'[24]

Specifically, in the case of non-binary ASE, my interest more directly is: 'I know I am not strictly a man or a woman. But this leaves me with a larger question that calls for aesthetic-expressive exploration and investigation: what *am* I? what does non-binary mean *in relation to me*?'.

§ 3. Barriers to Non-Binary Self-Expression

I now move on to establishing my second main claim: non-binary people experience barriers to aesthetic self-expression that prevent them from being able to gain its epistemic benefits.

---

[22] Alabanza, Travis (2023). *None of the Above: Life beyond the Binary*. Feminist Press at CUNY. See 15-34. This stereotype is deemed wrongful, partially, in that it attempts to frame a trans person's experience as affirming the dominant binary framework's conceptions of binary gender as constant, pre-established, and verifiable, when in truth, that framework excludes the significant gender innovation and investigations of trans individuals and communities.

[23] Revathi, *The Truth about Me: A Hijra Life Story*, 10. Emphasis added.

[24] Chandler, "*Art Has Been a Healing Tool*".

§ 3.1 The Dominant Framework as a Hostile Epistemic Environment

In this subsection, I argue that the dominant binary gender framework–the richly detailed and ubiquitous notion of gender being a matter of maleness and an oppositional femaleness–is epistemically hostile towards non-binary people.

As part of non-binary people's marginalisation as a social category, the collectively shared conceptual resources do not include them and their experiences[25]. Specifically, the dominant view of gender that informs most people's ASE is binary (male/female) and, due to a lack of hermeneutical participation on behalf of non-binary persons, does not acknowledge their experiences within and *beyond* its framework. Following the descriptions of hermeneutical marginalisation and injustice in 1, this consequent socio-epistemic position is one of hermeneutical marginalisation.

Recall that hermeneutical injustice involves a lacuna and narrative blanket that conceals the true nature of marginalised people's experiences. Non-binary people are hermeneutically marginalised insofar as their experience of gender is a lacuna in the binary gender framework. Where non-binary individuals do exist within the social imaginary, their identity is considered a fault of their own imagination. This perception of non-binary people paints non-binary individuals as not just false, but as dangerous and unpleasant: the collective understanding of the non-binary person is that of a threat to women's safety and feminist aims, mentally unhealthy and less competent than men or women, and even delusional, attention seeking, and the product of indoctrination[26]. This perception of non-binary people, I argue, acts as a narrative blanket that shields uptake and understanding of non-binary identities and gender frameworks, to non-binary people and binary-gendered people alike.

This marginalisation of non-binary people means that they have limited appropriate and helpful collective conceptual resources available to make their experiences intelligible (to themselves and to others)[27]. This results in a hostile epistemic environment from which non-binary people experience barriers to understanding. Therefore, the dominant gender framework produces a hostile epistemic environment for non-binary people, namely one that hermeneutically marginalises.

[25] Porkkala, *Hermeneutical Injustice, Nonbinary Gender Identities and Category Invalidation.*
[26] Fair Play For Women, (2021) "Non-Binary: The New Wolf in Sheep's Clothing Fair Play for Women." Fair Play for Women; McCarty, Megan, K, and Burt, Anna, H (2024) *Understanding Perceptions of Gender Non-Binary People: Consensual and Unique Stereotypes and Prejudice*, *Sex Roles*, 90; Murray, Douglas. 2022. "Non-Binary Is Short for 'Look at Me': Douglas Murray." Fox News. Many of these beliefs extend to the trans community as a whole. This suggests that a narrative blanket can extend beyond merely one subset of a group and perhaps even that these blankets might layer over one another in accordance with one's different identities.
[27] While a worthy consideration, I do not have space here to adequately discuss how hermeneutical marginalisation might dissipate in more accepting spaces. For discussion on this, please refer to Goetz's work on global and local hermeneutical economies in Goetze, Trystan S (2017) "*Hermeneutical* Dissent and the Species of Hermeneutical Injustice." *Hypatia* 33 (1): 73–90.

§ 3.2 Epistemically contaminated aesthetic self-expression (ECASE)

Following the discussion in section 2, that the dominant gender framework is a hostile epistemic environment to nonbinary people provides a great opportunity to use aesthetic self-expression as a cognitive tool to gain understanding in the face of a bad epistemic environment. In situations of epistemic hostility, ASE seems to be able to substitute for the typical epistemic routes to self-understanding and give us little signs as to who we are. So, when testimony is rejected and epistemic resources are scarce, ASE appears to be a route to finding glimmers of understanding through the conceptual dark glass, as Fricker puts it[28]. So, looking at A's case again, as a binary trans person, she appears to be in an environment characterised by typical epistemic hostility. Generally, her identity is ridiculed and overtly sexualised by those around her, meaning she can't rely on typical epistemic resources like talking her identity through with someone[29]. As such, the potential for epistemically beneficial conversations is limited and she is at a hermeneutical disadvantage.

Fortunately for A, however, the basis of A's gender identity as a woman is richly developed, including an aesthetic basis. This bank of epistemically loaded aesthetic conceptual material is seemingly what allowed for those precious aesthetic experiences to have their potent epistemic effects. What seems to have allowed for A's realisation of herself as a woman is a clear idea of what a woman tends to look like and seeing herself resemble this cognitive nest of aesthetic and social associations. Through ASE, those special moments where she was allowed to express herself lead to knowledge about herself; ASE was a valuable epistemic resource for her.

However, there seems to be two conditions under which the epistemic benefits of ASE are made less accessible: epistemic contamination of the aesthetic environment and epistemic contamination of relevant aesthetic objects. I refer to these under the umbrella of ECASE (Epistemic Contamination of Aesthetic Self-Expression).

Firstly, ASE can be prevented from eliciting those epistemic benefits when an exclusionary and/or false understanding is dominating the aesthetic playground. Seemingly, non-hostile aesthetic environments are only available as precious opportunities that allow for ASE when there are gaps or loopholes in the conceptual narratives about gender. So, for instance, in A's case, those revelatory experiences of ASE were only possible because of local festival shows that would allow for her to wear a costume, or in environments where people she knew from her hometown would not see her:

[28] Fricker, *Hermeneutical Injustice*, 148.
[29] Revathi, *The Truth about Me: A Hijra Life Story.*

> 'I insisted that I had [dressed up in the festival] for the goddess and so got away with it. On that occasion, my family decided not to make much of my dressing up and dancing.'[30]

> ''Why behave like this on a bus? Why apply make-up here? Look at how they're talking about us.' To which [the hijras whom she is travelling with] replied: 'Keep quiet! We can't be ourselves at home or do what we want to. There isn't anyone on this bus who knows us. *It's only like this, without anyone at home knowing, that can we do what we wish to do*—express our desire to be women.'[31]

So, while ASE can undermine the hostility of an epistemic environment through providing insight, it still is governed by that environment. This is the first kind of epistemic contamination – the pernicious conceptual load of an environment affecting someone's ability to conceptualise within it.

Secondly, the epistemic potential of ASE does not seem to be realised in cases where the epistemic hostility has leached into the aesthetic options themselves. In the case of non-binary ASE, the epistemic-aesthetic options within various performative art forms, sports, fashion, and so on, are regularly and overwhelmingly engendered and binary coded. In this we see epistemic contamination of aesthetic objects – the pernicious conceptual load of an environment affecting the aesthetic objects that lie within it, and thus, our ability to engage freely with them. Unfortunately, both kinds of epistemic contamination can be seen in the case of non-binary ASE, thereby hindering access to ASE's epistemic benefits.

The epistemic contamination of aesthetic environments and of objects are likely to overlap and influence one another, like in the following example:

> Aesthetic environments, like a hairdresser's wherein you can experiment with new self-expressive grooming styles, are overwhelmingly structurally epistemically contaminated. Finding a stylist who is willing to cut your hair to your liking is hard enough for most people, without also having to find someone who will cut your hair without misgendering you (especially if you're going for a style that is typical of your assumed sex), or will actually go for the gendered style you're requesting (if you're visually gender non-conforming). Additionally, there are gendered pricings whether you go to a barber or a hairdresser and added issues around finding a stylist who will work with your hair type. So, for some with black hair types wanting gender non-conforming styles, they can expect a much harder time reaping epistemic benefits from an emotionally, financially and stylistically risky experience than a cis white gender-conforming person might when getting their hair cut.

[30] Revathi, *The Truth about Me: A Hijra Life Story*, 10.
[31] Revathi, *The Truth about Me: A Hijra Life Story*, 15. Emphasis added.

The aesthetic options available when cutting your hair can also be difficult to navigate when hairstyles are gender coded. The pleasurable sensations of having long/short hair and the related aesthetic experiences possible with each (braiding, updos, etc), are to be considered alongside and potentially in conflict with the gendered look they are assigned to. Further, your desires for a style that represents you might have to be weighed up against what is considered 'appropriate' for your gender at your workplace and what will likely incur negative or positive treatment from others. It is thus, I argue, incredibly difficult to feel represented as a not-man-or-woman by aesthetic options that all carry the epistemic load of "man" or "woman".

This navigation between what a person wants in line with their gender and thus ASE and what is expected of their gender presentation brings into focus two features of epistemically contaminated aesthetic environments for non-binary people: the notion of androgyny and the threat of violence.

On top of those narratives surrounding non-binary people about mental illness and so on mentioned in 3.1, is a clear and flat conceptual image of what a non-binary person looks like. They have blue hair, are skinny, assigned female at birth (AFAB) and take on feminine social roles, and white[32]. When someone doesn't fit this stereotype, it's difficult to see themselves resemble the cognitive and social nest of associations of this identity. Thus, one is likely to struggle to have the same revelatory experience as discussed with A in 2.

Following the widespread internalisation of this perception is an expectation of performed androgyny. The notion of androgyny typically translates to a delicate balance of typical gendered features (e.g. makeup and a beard). Meeting this standard often means a tacit request to offset the gender coded body you live in, whether this be through what you wear (e.g. wearing more feminine clothes if you're assigned male at birth (AMAB), because trousers, while considered 'gender neutral' are not treated as such on an AMAB body) or adapting your frame (e.g. non-binary people who were assigned female at birth being expected to hide or even surgically modify their chest and hips). With this in view, even aesthetic options that, at first blush, seem epistemically uncontaminated, like supposedly gender-neutral or androgynous aesthetic figures and items, appear to also be contaminated by lingering binary conceptual frameworks. This culminates in a standard to the spirit of "you don't have to be seen as one gender or the other, just as long as you look like

[32] MacLean, Grace, (2026) *Of course you have blue hair and pronouns!*, Varsity Online; Bashforth, Emily. (2021). *The pressure to be thin as a non-binary person,* LGBTQ Nation; Klysing, Amanda, Marie Gustafsson Sendén, Emma Renström, and Anna Lindqvist (2023) "Gendered Stereotype Content for People with a Nonbinary Gender Identity." *Routledge Open Research* 2 (November): 45; Alabanza, *None of the Above: Life beyond the Binary*.

both in equal measure". This restricts the kind of autonomous expressive freedom ASE requires. So, even when non-binary people are interested in meeting this aesthetic goal in hopes of receiving appropriate gendered treatment in response to their presentations, this is difficult to do in an affirmative fashion.

Together, these considerations conceptually flood the aesthetic playing field and the epistemic hostility of that environment is exacerbated by literal hostility and aggression seen in anti-trans violence.

> 'Being visibly trans, consistently read as failing at either male or female gender, means I am lucky to walk outside and not be shouted at, or laughed at, or filmed within the space of a day.'[33]

When the epistemic contamination of aesthetic environments (for all trans people) includes a threat of violence, ASE isn't just difficult; it's dangerous. This fails ASE's safety requirement. As such, the epistemic contamination of non-binary ASE appears to limit ASE's revelatory possibilities.

A potential objection to this argument might be that non-binary people, in identifying as non-binary, reject the dominant binary framework. As such, the deconstruction of the dominant framework they've already done will likely bleed into their understanding and approach to gendered presentation. For instance, consider the following example:

> 'In an interview with The Washington Post, 18-year-old Kelsey Beckham, who identifies as nonbinary, states: "I don't want to be a girl wearing boy's clothes, nor do I want to be a girl who presents as a boy... I just want to be a person who is recognized as a person. That's how I'm most comfortable. *I'm just a person wearing people clothes*..."'[34]

Kelsey's testimony suggests that, in having deconstructed their view of gender enough to not identify with the strict binary categories, they have undergone a kind of epistemic de-contamination of self-expressive aesthetic options. Thus, this epistemic contamination might not prove such an issue for non-binary people when they have already undergone the necessary gender deconstruction to be unenthused by the binary gendered framework.

However, this quote precisely captures the problem I am talking about. In it, Kelsey expresses a desire to be recognised, not as a man or a woman, but as 'a person wearing people clothes.' While clothing may be just different cuts of fabric with a strange amass of social connotations, this conceptual weight is so readily applied to these clothes that it is hard not to feel like the clothes have these ideas weaved into them. That is, while we might think (or at least hope) that it is possible to

[33] Alabanza, *None of the Above: Life beyond the Binary*, 38.
[34] Dembroff, Robin, (2020) Beyond Binary: Genderqueer as Critical Gender Kind. *Philosopher's Imprint*, 20(9). See 7. Emphasis original.

epistemically decontaminate these aesthetic options, or be undeterred by these frameworks, it is problematic that non-binary people should have to undergo the presumably epistemically hefty task of decontaminating such widely and severely gendered aesthetic options and environments. For instance, when a clothing item's shape, colour, and texture is gendered, when you must go to separated and gendered floor of a shop to get it, when you must pick a gendered changing room to try it on in (if you're permitted to choose), deal with people staring, pointing, laughing, recording and shouting abuse at you, and when you're addressed by 'sir' or 'ma'am' when trying to buy it from the cashier, the fun of trying on something, attempting to feel out how it makes you feel, and trying to get closer to the answer of 'does this feel like me? Is this reflective of who I am?' has been heavily tampered with. ASE is no longer a viable cognitive tool, and as such, your ability to make sense of and interpret yourself and your experiences freely and playfully has been negatively impacted. In this, we see how other factors like physical manifestations of the gendered social imaginary such as sexed changing rooms can impact one's ability to access certain cognitive tools.

So, while someone might know that they are not (exclusively) a man or a woman, the dominant binary framework persistently rises to the surface and epistemically contaminates their perception of these clothes as such. One's ability to have a gender-neutral understanding of their clothes is impeded by the constant reminders that (what feels like) the rest of the world disagrees with them. One's engagement with their clothes is both a personal engagement with the self and one's identity, and a way in which others come to understand and treat that person. This is a particularly pressing consideration when keeping in mind the literal threat of violence faced by non-binary people wishing to express gender non-conformity in many aesthetic environments. As such, I suggest that one's ability to use their aesthetic expression through things like clothing, is being undermined.

With this discussion in mind, we can define ECASE as follows:

> Someone is a victim of ECASE when they experience epistemically contaminated aesthetic self-expression that (a) results in significant cognitive disablement of their understanding of themselves and/or their experiences *and* (b) is the product of an unjust epistemically hostile environment (typically, one of hermeneutical marginalisation).

(a) follows Fricker's explanations of how gaps in shared conceptual resources affect people differently according to how they bear relevancy to the person's 'cognitive disablement.'[35] That is, just because many aesthetic self-expressive options are epistemically contaminated, this shared socio-epistemic environment does not mean ECASE is experienced by all groups.

For instance, while we're all working with the same conceptual resources about gender, the hermeneutic barriers cis (and to a degree, given that the dominant binary

[35] Fricker, *Hermeneutical Injustice*, 153.

conception doesn't include their transition and bodily differences, binary trans) people experience do not typically seem to resemble the severity of those felt by non-binary ECASE victims. Cis people, by virtue of identifying with the binary enough to call themselves men or women, must possess a significant level of self-representation within that framework to do so. As such, this collective understanding of what it means to be a man or a woman is a fruitful collective resource to tap into when finding out who they are, even if this means doing the work of picking and choosing which gendered concepts to opt out of. So, the conceptual load on the available aesthetic options is going to favour their self-understanding, despite some potential areas of difficulty. For instance, when a woman is trying to find out what 'woman' means for her, this might involve difficulties in ASE when encountering epistemically loaded aesthetic materials such as makeup.[36] For example, feeling that you have to use makeup to be treated well as a woman impedes one's ability to engage with it freely and have full creative control over how/whether one applies it. Though this may be a hiccup in trying to figure out what being a woman for her is, a minor instance where an otherwise richly developed, representational, and supportive conceptual framework presents difficulties, this doesn't look like ECASE as it does not meet the first criteria (a), resulting in significant cognitive disablement of their experiential and/or self-understanding. I suspect that this is true of the gendered experience of many cisgender heterosexual people. However, where we might think marginalisation tracks across multiple social categories, or where this epistemic hostility is very widespread, we might see these cognitive effects become significant enough to be instances of ECASE (as we will see in the suggested cases in 3.2).

This significant cognitive disablement also needs to be the product of unjust epistemic hostility to be a case of ECASE. So, someone who has lost their hair due to cancer treatment is likely to experience significant epistemic consequences, with their aesthetic options regarding their hair being forcefully changed by their circumstances. Their self-understanding might be difficult to parse given this significant change. However, this is not a case of ECASE, because this isn't due to unequal hermeneutical participation or widespread unjust epistemic hostility; it is just epistemic misfortune.

In this, (b) captures the unfairness expressed in both the expectation of androgyny and the threat of violence non-binary people face. These aren't the products of epistemic misfortune, rather, the epistemic-aesthetic harm of not being able to understand oneself in this case is due to the unnecessary epistemic hostility of the current binary framework. The engenderment of clothing, hairstyles, colours, shapes, sections of department stores, bodily figures resembles an invented social framework. The hermeneutical marginalisation underpinning this structure, as

[36] Cahill, Anne, J (2003) "Feminist Pleasure and Feminine Beautification." *Hypatia* 18 (4): 42–64; Jeffreys, S, (2014) *Beauty and Misogyny: Harmful cultural practices in the West* (2nd ed.), Routledge.

detailed in 1 and 3.1, is what frames the unfairness of this epistemic circumstance experienced by non-binary people.

So, where an unjust epistemically hostile environment results in severe epistemic contamination of someone's ASE, we can consider them a victim of ECASE. Thus, in its significant epistemic disablement and its context of hermeneutical marginalisation, navigating non-binary self-expression is an example of ECASE. This notion of injustice invites us to consider ECASE as it relates to hermeneutical injustice.

§ 4. Barriers to Non-Binary Self-Understanding: A New Form of Hermeneutical Injustice

Hermeneutical injustice can be enacted through atypical means, namely, aesthetic ones. ECASE possesses the following features of hermeneutical injustice: hermeneutical marginalisation, unavailable conceptual resources, unfairness, and a clear potential to impact one's social experiences. However, I want to characterise ECASE as a *form* of hermeneutical injustice, rather than an *instance* of cognitive hermeneutical injustice or an exclusively aesthetic *form* of injustice.

Having argued for ECASE's cognitive harm within the typical hermeneutical framework, I maintain that exclusively focusing on traditional hermeneutical framework misses the distinct role of the aesthetic in the hermeneutical injustice under consideration. It is not merely that the hermeneutical marginalisation of non-binary people makes them struggle to understand themselves and that this later manifests in aesthetic-epistemic difficulties. Instead, it appears to exist as a means in which the original hermeneutical injustice occurs; the method of understanding, the conceptual tool of ASE, becomes more difficult to use; it is a constitutive part of that hermeneutical injustice.

Further, I argue against strictly categorising this phenomenon as an aesthetic injustice. Alongside the clear epistemic role, I feel that this would also be an inappropriate use of 'aesthetic injustice' which, in my view, is the unjust hindering of one's aesthetic/affective mental tools – such as one's ability to appreciate art, have specific aesthetic sensations, and such[37]. I want to justify this distinction further, by suggesting that this is a separate phenomenon that is also likely to happen in the case under consideration. That is, when one's aesthetic options are epistemically contaminated and one's body is forcibly given non-representational meaning, I suggest one is likely to be unable to aesthetically appreciate those items and bodily features, a separate but interesting phenomenon to the hermeneutic role of ASE/ECASE. So, to have an accurate picture of what is going on in these instances of epistemic contamination, I want to suggest this separation so that further cases under consideration can be addressed with the appropriate diagnostic conceptual

[37] See Dalaqua, Gustavo, H. (2020) Aesthetic injustice, *Journal of Aesthetics & Culture*, 12(1), 171-183.

tools. So, the case of nonbinary ECASE stands as not merely a unique *instance* of hermeneutical injustice but relates to a specific *form* of hermeneutical injustice. Furthermore, I argue that the threat of ECASE extends to multiple groups experiencing hermeneutical marginalisation, not just non-binary people. Further applications of ECASE may include:

(i) In a western anti-black epistemic environment, black people are likely to experience traditional hermeneutical injustice, by virtue of the epistemic hostility and likely hermeneutical marginalisation of that environment. As a result, black people may struggle with understanding themselves within a social system that tells them they are reducible to racist stereotypes[38]. Additionally, if they internalise the racist propaganda of that environment, they may experience aesthetic injustice, in that they may struggle to aesthetically appreciate art that makes intentional and overt references to black history and subcultures[39]. Furthermore, hermeneutical injustice may also rear its head via ECASE; when whiteness heavily informs the western standard for what is deemed attractive, professional, and aesthetically valuable, this epistemic contamination of aesthetic options may lead a black person to struggle to gain the epistemic benefits of ASE when trying to understand where they exist in relation to the social groups they are a member of.

(ii) In *Old Age*, Beauvoir describes feeling disconnected from other's perspective of her age[40]. While she understood herself as being the same sort of person she was decades before, she describes being continually met with others treating her differently for being old. The aesthetics of aging are laden with epistemic cruelties; shared conceptions of old people being disposable, dim-witted and objects of disgust, particularly in sexual contexts[41]. Beauvoir describes how trying to understand herself with these tampered aesthetic-epistemic conceptual tools was difficult to the point of being unattainable, with her claiming that to be this old woman that others perceived her as was an unrealisable, an impossible task[42]. This again looks like a case where ECASE has affected her ability to understand herself and her experience of aging.

If applicable, these cases suggest that non-binary ECASE is not merely a unique case or side-effect of hermeneutical injustice but instead strengthens the claim that ECASE is a *form* of hermeneutical injustice.

---

[38] Hull, G. (2016) Black Consciousness as Overcoming Hermeneutical Injustice, *Journal of Applied Philosophy*, 34(4), 573–592.
[39] Dalaqua, *Aesthetic injustice*, 171-183.
[40] Beauvoir, Simone de. (1977) *Old Age*, Harmondsworth, Middlesex: Penguin Books.
[41] See Beauvoir, *Old Age*, 542 & 300-301.
[42] See Beauvoir, *Old Age*, 291.

§ 5. Conclusion

In summary, I have argued that ECASE is a new form of hermeneutical injustice, with non-binary ECASE being my core empirical case. I have argued this by first claiming that aesthetic self-expression is an epistemically valuable experience that can give someone glimmers of understanding when in a position of epistemic hostility. I have subsequently claimed that non-binary people experience hermeneutical marginalisation but that the epistemic contamination of their aesthetic self-expression means that they are unable to gain these epistemic benefits, unlike binary transgender people. Following this, I have argued that this epistemic contamination of aesthetic self-expression is a novel form of hermeneutical injustice.

This conclusion suggests that hermeneutical injustice elicits its harm through more ways than previously considered, and as such, suggests that we must be attentive to the possibilities of more widespread ECASE and other aesthetic-epistemic injustices.

Further lines of inquiry beyond the scope of my discussion here may lead to whether the affective has a similar epistemic contamination problem and whether that is another kind of injustice, how epistemic contamination (and other potential aesthetic-epistemic injustices) relate to impeded abilities to understand things other than the self, and perhaps most crucially, what good epistemic decontamination looks like and how we can achieve it.

References

Alabanza, Travis (2023) *None of the Above: Life beyond the Binary*. Feminist Press at CUNY.

Bashforth, Emily (2021) 'The Pressure to Be Thin as a Non-Binary Person - LGBTQ Nation', LGBTQ Nation. https://www.lgbtqnation.com/2021/11/pressure-thin-non-binary-person/.

Beauvoir, Simone de (1977) *Old Age*. Penguin Books.

Brownmiller, Susan (1990) *In Our Time: Memoir of a Revolution*. Dial Press.

Cahill, Ann J (2003), 'Feminist Pleasure and Feminine Beautification', *Hypatia* **18**: 42–64. DOI: 10.1353/hyp.2003.0073

Carmona, Carla (2022) 'Binarism Grammatical Lacuna as an Ensemble of Diverse Epistemic Injustices', *Social Epistemology* **37**: 339–63. DOI: 10.1080/02691728.2022.2103473

Chandler, Katie (2019) 'Art Has Been a Healing Tool', Mind https://www.mind.org.uk/information-support/your-stories/art-has-been-a-healing-tool

Clanchy, Nick (2024) 'Tackling Hermeneutical Injustices in Gender-Affirming Healthcare', *Hypatia* **39**: 688–710. DOI: 10.1017/hyp.2024.15

Dalaqua, G H (2020) 'Aesthetic Injustice', *Journal of Aesthetics & Culture* **12**: 171–83. DOI: 10.1080/20004214.2020.1712183

Fair Play For Women (2021) 'Non-Binary: The New Wolf in Sheep's Clothing Fair Play for Women', Fair Play for Women. July 11, 2021. https://fairplayforwomen.com/non-binary-the-new-wolf-in-sheeps-clothing/

Fricker, Miranda (2007) 'Hermeneutical Injustice', in Epistemic Injustice: Power and the Ethics of Knowing. Oxford: Oxford University Press.

Goetze, Trystan S. 2017. 'Hermeneutical Dissent and the Species of Hermeneutical Injustice', *Hypatia* **33**: 73–90. DOI: 10.1111/hypa.12384

Hull, George (2016) 'Black Consciousness as Overcoming Hermeneutical Injustice', Journal of Applied Philosophy **34**: 573–92. DOI: 10.1111/japp.12201.

Jeffreys, Sheila (2014) *Beauty and Misogyny*. Routledge. DOI: 10.4324/9781315771458

Klysing, Amanda, Marie Gustafsson Sendén, Emma Renström, and Anna Lindqvist (2023) 'Gendered Stereotype Content for People with a Nonbinary Gender Identity', Routledge Open Research 2 (November): 45. DOI: 10.12688/routledgeopenres.17976.1

MacLean, Grace (2026) 'Of Course You Have Blue Hair and Pronouns!', Varsity Online. https://www.varsity.co.uk/fashion/29568

McCarty, Megan K, and Anna H Burt (2024) 'Understanding Perceptions of Gender Non-Binary People: Consensual and Unique Stereotypes and Prejudice', *Sex Roles* **90**: 392-416. DOI: 10.1007/s11199-024-01449-2

Murray, Douglas (2022) 'Non-Binary Is Short for 'Look at Me': Douglas Murray', Fox News. https://www.foxnews.com/video/6312447412112

Porkkala, Siiri (2024) 'Hermeneutical Injustice, Nonbinary Gender Identities and Category Invalidation', *Social Epistemology* **39**: 403–13. DOI: 10.1080/02691728.2024.2403625

Revathi, A, and V Geetha (2010) *The Truth about Me : A Hijra Life Story*. Penguin Books.

Rosola, Martina (2024) 'Linguistic Hermeneutical Injustice', *Social Epistemology* **39**: 425–35. DOI: 10.1080/02691728.2024.2401143

Undergraduate Philosophy Journal of Australasia www.upja.online
Vol. 8, August 2026

Can F. P. Ramsey's Theory of Belief Accommodate Akrasia?

Matthew DeKock[1]

University of Toronto

Abstract

Cheryl Misak has expressed concern that Frank Ramsey's dispositionalist theory of belief may have difficulty accommodating *akrasia*—the phenomenon of agents intentionally acting against their better judgement in the face of countervailing pleasures and/or pains. She also mentions that Ramsey might be able to handle *akrasia* with recourse to his theory of partial belief and desire. This paper develops this position, arguing that Ramsey can indeed accommodate *akrasia*. I demonstrate that Ramsey's theory of partial beliefs and his treatment of humans as expected utility maximizers allows for cases where desire overwhelms our normative beliefs, hence producing acts of *akrasia*. The paper concludes with a sketch of how this account can be extended to account for *enkrateia* and virtue, as well as an analysis of how well the Ramseyian treatment I put forth captures the original meaning of *akrasia*.

[1] *Matthew DeKock is an undergraduate student at the University of Toronto where he is pursuing a specialist degree in philosophy. His research interests include pragmatism, ancient philosophy, epistemology, and ethics.*

§ 0. Introduction

A difficulty for any action-oriented epistemology is whether it can make sense of *akrasia*, or 'weakness of the will,' the phenomenon of acting against one's better judgement (i.e. against one's normative beliefs) in the face of countervailing pleasures and/or pains. For frameworks which identify beliefs with actions, it seems that this phenomenon will be impossible to accommodate—as what it is to believe a normative proposition is to act according to it. Of course, this is an absurd result as likely everyone reading this has experienced weakness of the will at some point throughout their life. Frank Ramsey's account of belief is one such action-oriented theory, concerned with the dispositions of agents, and so *akrasia* needs explaining under his framework. I will argue that Ramsey can indeed make sense of *akrasia* with recourse to his theory of how our desires interact with our partial beliefs to determine action. In section I, I will exposit the relevant theoretical background on *akrasia* and Ramsey's theory of belief, and set up the potential problem with combining them. In section II, I will develop my thesis that Ramsey can accommodate *akrasia* by putting forth a definition of it in terms of partial belief, desires, and expected utility. Finally, in section III, I will sketch how my analysis can further be extended to account for *enkrateia* (or 'strength of will') and virtue, as well as discuss the extent to which my Ramseyian analysis of *akrasia* still captures its original meaning.

It should be noted that this paper sits squarely within the pragmatist tradition, on which one of the primary virtues of a theory is its success in guiding action. Accordingly, several of the argumentative moves that I make rely on the pragmatist principle that a theory's practical merit is what chiefly justifies accepting it or rejecting it.

§ 1.

Akrasia is the phenomenon of intentionally acting against your normative beliefs in the face of strong desires 'pulling' in the opposite direction. In English, it means something like 'lack of self-control' or 'weakness of the will'. On Aristotle's account, *akrasia* is a negative state of character and, in becoming virtuous, an agent passes from *akrasia*—when their beliefs about the good are correct, but their desires are misaligned, and they follow their desires rather than what is right; to *enkrateia*—when their desires are still misaligned with their beliefs but they learn to choose the right over the pleasurable; to virtue—when their desires finally align with their (true) normative beliefs about the good.

Socrates' position on the matter was that, because the nature of normative belief is inherently motivational, *akrasia* does not really exist because one cannot believe a normative proposition without acting according to it.[2] It should be clear

[2] Aristotle. 2020. Nicomachean Ethics. Translated by Adam Beresford. London: Penguin Books, 1145b25-30.

that this account is untenable because it would force us to deny our regular experience of *akrasia*, and this sort of highly impractical doubt of our experience is something we are unwilling to commit to. As C.S. Peirce would put it, doubting our experience of *akrasia* is not a "real and living doubt, and without this all discussion is idle." [3] It is impossible for this doubt in the existence of *akrasia* to grip us, and so becomes essentially a 'paper doubt'.

Aristotle recognized this trouble, and made an attempt to get around it by affirming the existence of *akrasia*, and treating it as a case in which someone has knowledge of the good but does not use it.[4] Whether Aristotle's account is sufficient on his own epistemology is an open question and far outside the purview of this paper. However, it should be clear that this account will not work for Ramsey's theory of belief because, for Ramsey, believing something is essentially to have particular dispositions to act in a certain way, so the idea of knowledge that is divorced from usage is nonsensical.

Nonetheless, the Socratic denial of akrasia and the Aristotelian response to it constitutes a philosophical legacy that leaves us with which any theory of belief must meet when it comes to the phenomenon of *akrasia*. The first, as recognized by Aristotle, is that any theory of belief must be able to accommodate *akrasia*. The second, which led Socrates to deny *akrasia* in the first place (as he believed the following condition could not be met): in building *akrasia* into our epistemology we have to somehow retain the motivational nature of normative belief.

1.1

Ramsey's is a causal theory of belief, meaning that belief is defined in terms of its causes and effects. Particularly the behavioral effects of belief are what is relevant to its analysis—that "we deduce from [the belief] and act on it in a certain way."[5] Of course, we hold plenty of beliefs that we are not at any given moment acting on, like my belief that Caesar is dead. So, instead of a belief being something "which does actually lead to action," Ramsey asserts that belief is something "which would lead to action in suitable circumstances."[6] For example, my belief that Robarts Library at the University of Toronto is located on Huron street amounts to my tendency to, when looking to study at the library, head up Huron street and expect to reach it. As such, beliefs are dispositional states, defined as ways in which agents *would* act, *if* they were placed in the proper circumstances. One of the examples Ramsey himself gives of his analysis is the belief that Hell exists: "To say a man believes in hell means … that he avoids doing those things which would result in his being thrown

[3] Peirce, Charles S. 1877. "The Fixation of Belief." Popular Science Monthly 12 (November): 1—15.
[4] Aristotle, Nicomachean Ethics, 1146b30-35.
[5] Ramsey, Frank. (1929a/1990) 'General Propositions and Causality' in Philosophical Papers. Ed. D. H. Mellor. Cambridge: Cambridge University Press, 145-164, 159.
[6] Ramsey, Frank. (1926/1990) 'Truth and Probability' in Philosophical Papers. Ed. D. H. Mellor. Cambridge: Cambridge University Press 52-95, 66.

into hell."[7]. The behaviors he is speaking of are fairly obvious, including practically anything[8] contrary to the Bible (so long as the agent views the Bible as an adequate guide to the truth).

It should be immediately clear that the notion of *akrasia* discussed in the previous section (I.i) seems to directly contradict Ramsey's theory of belief. As Cheryl Misak points out, a belief "about the good expresses a disposition to act, [and so] it seems as though it will never be the case that I believe 'φ-ing is the best thing to do, all things considered' and yet find myself more strongly disposed to refrain from φ-ing."[9] To put it another way, it seems as though the fact that (on Ramsey's theory of belief) one cannot hold a belief that is contradictory to their dispositional states, one can never fail to act according to their belief. To defend Ramsey from this objection, more of his theory of belief must be introduced. The remainder of this section (I.ii) will serve that purpose.

Ramsey's theory of belief admits of partial beliefs, one of his most valuable contributions to philosophy. He says that "the degree of belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it."[10] The measurement of belief begins by asserting that full belief can be denoted "by 1, full belief in the contradictory by 0, and equal beliefs in the proposition and its contradictory by ½."[11] As for the other values not yet discussed between 0 and 1, Ramsey proposes that we can measure these degrees if we:

> [T]ake as a basis a general psychological theory, which is now universally discarded, but nevertheless comes, I think, fairly close to the truth…. I mean the theory that we act in the way we think most likely to realize the objects of our desires, so that a person's actions are completely determined by his desires and opinions. [12]

His reasoning for taking this as given that we are essentially expected utility[13] maximizers, even though this psychological theory is widely considered to be wrong, is that it, like Newtonian Mechanics, can be used to approximate the correct answer quite closely. Given Ramsey's commitment to pragmatism—the metaphilosophical position on which a theory is evaluated by its practical success in guiding action and/or inquiry—and thereby some sort of fallibilism, it should not be surprising that he is willing to approximate in his theory, even knowing it is not perfectly accurate, rather than simply be unwilling to propose a theory at all.

This same pragmatist/dispositionalist framework lends Ramsey's theory of belief one of its most distinctive strengths: its ability to handle the phenomenon that

[7] Ramsey, Frank (1930/1991). 'On Truth'. Ed. N. Rescher and U. Majer. Dordrecht: Kluwer. 91.
[8] Including, perhaps, my behavior of capitalizing 'Hell' throughout this paper.
[9] Misak, Cheryl. 2018. *Cambridge Pragmatism*. Oxford University Press. 219.
[10] Ramsey, *Truth and Probability*, 65.
[11] Ramsey, *Truth and Probability*, 64.
[12] Ramsey, *Truth and Probability*, 69.
[13] Not in the sense of pleasure, but in the sense of a 'desire' (before the fact), which may or may not be pleasurable (after the fact).

"no one feels strongly about things he takes for granted," or the fact that our most dearly held beliefs are accompanied by (practically[14]) no 'belief-feelings'.[15] An example which makes use of this strength is my belief that the ground is solid—a belief I hold almost entirely without conscious consideration. The intuitive explanation or evidence for this particular belief is dispositional: that I tend to walk on the floor, to not expect it to swallow me up, to express skepticism when someone asserts the converse, and so on. On a traditional Humean[16] view of belief a particular belief is constituted by various belief-feelings. When confronted with a belief that is expressed dispositionally and which is absent[17] of belief-feelings, such as the belief that the ground is solid, a Humean must either deny that the subject really has the belief in question or deny their own approach. Ramsey's theory, by contrast, is not faced with this unfortunate disjunction.

Still, Ramsey's theory of belief does have some counterintuitive consequences, particularly that it forbids beliefs being evaluated in terms of their truth and falsity. It would be odd to say, for example, that my tendency to avoid doing things that would damn me to Hell is 'true'. It can be true or false *that* I am disposed in certain ways and it can be true or false *that* Hell exists, but the dispositions themselves cannot be.[18] The alternative for Ramsey is that the analysis of belief falls to success or failure; beliefs are to be analyzed like "a map… by which we steer."[19] If the map is good, following it will get us to our destination. It should be made clear that Ramsey's account does not permit for beliefs being 'good' in absence of the propositional content therein being true—viz. a belief in p is only useful if p.[20] Now having laid out the core of Ramsey's theory of belief, we will again take up the problem of *akrasia*.

Misak also notes that Ramsey might be able to accommodate *akrasia* with reference to his account of how beliefs combine with desires.[21] In the next section (II) I will develop the thesis that Ramsey can indeed accommodate *akrasia* with recourse to his theory of partial beliefs and his theory on how desires combine with these beliefs to determine our actions.

[14] Beyond the conscious experience of considering it to be an example of an unconscious belief.
[15] Ramsey, *Truth and Probability*, 65
[16] Hume, David. 1739. *A Treatise of Human Nature*. Oxford University Press. Book 1, Part 3, §7
[17] Not wholly absent, but I certainly believed that the ground was solid when I walked to, for example, Robarts Library this afternoon, and I certainly did not have any belief-feelings whilst doing so.
[18] Ramsey does assert that there are 'other cognitive attitudes' relevant to belief in *General Propositions and Causality* (Ramsey 1929a, 147-148) but it's clear that, whatever these may be, they are not simply truth and falsity.
[19] Ramsey, *General Propositions and Causality*, 146.
[20] Ramsey, Frank. (1927/1990) 'Facts and Propositions' in Philosophical Papers. Ed. D. H. Mellor. Cambridge: Cambridge University Press, 34—51, 40.
[21] Misak, *Cambridge Pragmatism*, 219.

§ **2.**

I will adopt the following definitions for the remainder of this essay:

- W{1} = That {1} is the wrong thing to do, all things considered
- p = Some proposition
- The function which expresses the subject's degree of belief in p as a number on the closed interval [0,1] shall be denoted as B(p)
    - As for the degree of belief in the contradictory, this can be represented as 1 — B(p)
- The function which expresses the subject's desire for a certain state of affairs (which includes p) as an integer on the open interval (-∞,∞) shall be denoted as D(ω_1 & p) [22]
    - As for the desire for the reverse, this shall be denoted as D(ω_2 & ¬p)
- g_X = the gamble 'X' [23]
- The function expressing the expected utility of a given (binary) gamble X on the truth of p shall be denoted as EU(g_X), and its equation is given as follows:

$$EU(g_X) = B(p) \bullet D(\omega_1 \,\&\, p) + (1 — B(p)) \bullet D(\omega_2 \,\&\, \neg p)$$ [24]

- When the EU(g_X) > EU(¬g_X), the subject is determined[25] to gamble, and vice versa.

The account of *akrasia* I wish to advance is that when a subject takes a specific gamble on the truth of W(x), which I shall refer to as a 'Gamble of Indulgence', the agent acts out of *akrasia*. More specifically, taking a Gamble of Indulgence is *identical* to acting out of *akrasia*.

A Gamble of Indulgence (g_I) is defined as follows (i.e. each of the following conditions is necessary and they are jointly sufficient to consider any gamble a Gamble of Indulgence):

1. The subject has a high value for B(W(x))

    (i.e. the subject has a high degree of belief that x is wrong)

[22] It should be noted here that I am using omega, ω, to represent a possible world.

[23] Particular gambles designated with a name shall replace the subscript of X with the first letter of the name of that specific gamble

[24] See: MacBride, Fraser, Mathieu Marion, Frápolli, María José, Dorothy Edgington, Edward Elliott, Sebastian Lutz, and Jeffrey Paris. 2019. "Frank Ramsey." Stanford.edu. August 14, 2019. https://plato.stanford.edu/entries/ramsey.

[25] See Ramsey in *Truth and Probability*, quoted on Page 4 of my paper.

2. The subject has a negative, low magnitude, value for $D(\omega_1 \ \& \ W(x))$

   (i.e. the subject disprefers doing x when W(x), but not to a large extent)

3. The subject has a positive, high magnitude, value for $D(\omega_2 \ \& \ \neg W(x))$

   (i.e. the subject prefers doing x when ¬W(x), to a large extent)

As opposed to motivating this account one condition at a time, I will propose an example which I believe to be illustrative of my account, that of 'The Strawberry Cheesecake Lover':

> The Strawberry Cheesecake Lover is a subject who believes quite strongly that he should not indulge himself in unhealthy foods, as he is quite overweight and trying to become healthier. His favorite food in the whole world is strawberry cheesecake and his desire for it is at an all time high since he is on a diet. Our subject takes a stroll through the park and is greeted by a friend offering him a slice of strawberry cheesecake. He fights dearly to retain his principle of temperance, but fails and indulges in the cake.

Assume that:

a = the act of indulging in the cake

- $B(W(a)) = 0.8$
  - Assigned as an arbitrarily high value to account for his high credence in the normative proposition W(a).[26]
- $D(\omega_1 \ \& \ W(a)) = -10$
  - Assigned as a negative and relatively low value (magnitude) to account for the fact that this is only a single instance of imprudence in the case that W(a), and our subject cares less about the individual instances and more about the overall virtue.
- $D(\omega_2 \ \& \ \neg W(a)) = 100$
  - Assigned as a positive and high value (magnitude) to account for the fact that this is his favorite food and that he should very much like to eat it if it's not wrong to.

(Note that the points above are sufficient to consider eating the cake a Gamble of Indulgence)

- $EU(\neg g_I) = 10$

[26] Because this is a thought experiment, these specific values are stipulated for ease of illustration. In real-world cases, however, Ramsey's theory will—by observing an agent across a range of different situations—generate enough data to calculate both $B(W(a))$ and $D(\omega_1 \ \& \ W(a))$ (and their complements). In fact (on my reading), the entire reason Ramsey treats us as expected utility maximizers is to be able to reliably deduce these values given repeated and intentional observation. See Truth and Probability (page 68ff) for Ramsey's comments on this.

- Assigned as a positive and relatively low value (magnitude) to account for the fact that this is only a single instance of abstinence, and, as previously mentioned, our subject cares less about the individual instances and more about the overall virtue.

Plugging these values into the following equation and solving for $EU(g_I)$:

$EU(g_I) = B(W(a)) \bullet D(\omega_1 \,\&\, W(a)) + (1 — B(W(a))) \bullet D(\omega_2 \,\&\, \neg W(a))$

We get that $EU(g_I) = 12$, and we had stipulated that $EU(\neg g_I) = 10$. Hence the expected utility of gambling is higher than that of not ($EU(g_I) > EU(\neg g_I)$), and the subject takes[27] the Gamble of Indulgence, thereby falling into *akrasia*.

It should be clear that this Ramseyian treatment meets the two criteria I posited in section I, I.i. Regarding the first condition, *akrasia* can be handled by Ramsey as I have described above in this section, with recourse to his theory of partial beliefs and how desires combine with them to determine action. As for the second, the motivational nature of normative belief has been retained because it is the nature of any belief to be 'motivational'—meaning that they contribute to motivating our actions—as beliefs are, in essence, dispositional states.

§ 3.

Any theory of *akrasia* also must be capable of being extended to account for *enkrateia*, or 'strength of will,' as well as virtue. I will sketch how Ramsey may be able to handle these but refrain from discussing them at length. My proposal for how Ramsey could handle these is as follows: an *enkrateic* act occurs when an agent is presented with a Gamble of Indulgence but does not gamble; a state of virtue is a state of never being presented with a Gamble of Indulgence because the agent's desires align with their (correct) beliefs about the good.

Aristotle defines *enkrateia* as a state wherein one's desires are misaligned with the good, the same as the *akratic*, but unlike the *akratic* they overcome these desires and act rightly.[28] Because Ramsey is treating people as expected utility maximizers, not gambling when presented with a potential Gamble of Indulgence would mean that the expected utility of gambling must be lower than that of not. There are two ways of achieving this, either lowering the value for the misaligned desire—$D(\omega_2 \,\&\, \neg W(x))$—or raising the degree of belief that the act is wrong—$B(W(x))$. Admitting Aristotle's condition that the desires of the *akratic* and the *enkrateic* are the same, the first option is not available, so the degree of belief must increase. In the case of the Strawberry Cheesecake Lover, their degree of belief that eating the cake is wrong, i.e.

[27] It should be noted once again that, in treating humans as expected utility maximizers, whenever the expected utility of a specific gamble exceeds that of not gambling (as it does here), the subject is *determined* to gamble.
[28] Aristotle, *Nicomachean Ethics*, 1145b10-15.

B(W(a)), would need to be >0.82 (roughly) to produce a $EU(g_I) < EU(\neg g_I)$ so that the subject would act from *enkrateia*.

On virtue, Aristotle refers to it as the state of having one's desires aligned with the good, unlike the *akratic* or the *enkrateic*, so they do not need to exercise self-control.[29] On my Ramseyian interpretation, this would mean that there is never a case in which a subject's $D(\omega_2 \,\&\, \neg W(x))$ is positive and high magnitude[30] at the same time as their B(W(x)) approaches 1 and is correct.[31] This stipulation, of course, violates the conjunction of necessary conditions (1) and (3) for considering a gamble to be a Gamble of Indulgence. Therefore, an agent being virtuous, on this view, amounts to them never being presented with a Gamble of Indulgence due to their desires never being misaligned with their (correct) beliefs.

3.1

The question now before us is whether the account I have put forth accurately captures the same *akrasia* that Aristotle is referring to—what I take to be the original meaning of the term. One portion of this original account that I have neglected to discuss is that Aristotle assumes that the *akratic* is "100 percent certain" of W(x).[32]

It should be immediately obvious that my account does not admit of this possibility. To have a belief of degree 1 in W(x) (and therefore 0 in the contradictory) implies that, whenever presented with a Gamble of Indulgence, the subject will refrain from gambling, even if the desire for $D(\omega_1 \,\&\, \neg W(x))$ approaches infinity. Hence, in treating humans as expected utility maximizers, we actually forbid what appears to be the original account of *akrasia*, and seem to fall directly into the kind of Socratic denial that we set aside as a paper doubt in section I.i.

The response, I think, is to deny that Aristotle can really support the idea that an *akratic* is 100% certain of their normative beliefs. If he cannot, then my account aligns his—i.e. the original meaning of the concept—in this regard. All I can say on this matter is to assert a pragmatist and/or fallibilist principle that no human person ever really reaches a degree of belief of 1 (or 0) in any proposition whatsoever, and so the real human phenomenon of *akrasia* cannot rely on us having a degree of belief in W(x) of 1. If it did, then Aristotle would also fall into the kind of Socratic denial that he agrees to be absurd.

Ramsey would accept this principle, as he in fact states it outright:

---

[29] Aristotle, *Nicomachean Ethics*, 1104b1-5.

[30] Whether it may be positive as long as it is low enough magnitude or must be negative is a topic for another paper.

[31] Of course, because beliefs are not truth-apt for Ramsey, correct means something like 'the belief is useful'. Then again, because Ramsey asserts that a belief in W(x) cannot ever be useful if ¬W(x), whenever W(x) is false, this condition of 'correctness' cannot obtain. Whether the truth of W(x) or the usefulness of the belief in W(x) matters more for virtue is also a topic for another paper.

[32] Aristotle, *Nicomachean Ethics*, 1146b30-35.

> Ideally [the] best thing is that we should have beliefs of degree 1 in all true propositions and beliefs of degree 0 in all false propositions. But this is too high a standard to expect of mortal men.[33]

I think the apportioning of certainty to the *akratic* comes from the more typical Humean view of belief that a belief is constituted by certain feelings. Because *akrasia* comes with incredibly strong belief feelings, we propose certainty. However, this misses a key observation Ramsey makes that "no one feels strongly about things he takes for granted," and so the presence of such intense belief feelings should actually be interpreted as further evidence that there is a *lack* of certainty when dealing with *akrasia*.

It remains to be argued that certainty is unachievable, and it needs to be assumed in order to avoid my account failing to capture the original meaning of *akrasia*. Then again, even if it turns out to be that there are some normative[34] propositions of which we can be certain, it still needs to be shown that Aristotle's certainty condition is correct. It seems to me that that it is clearly not, as the 'fix' for the *akratic* that we tend to apply is to try and *convince them further* that their act is wrong. If they were already certain, then there would be no point in trying to bring them closer to certainty. Regardless, these are topics for further literature to address.

§ 4. Conclusion

In this paper I have argued that Ramsey's dispositional theory of belief can indeed be extended to account for *akrasia* and that this treatment also succeeds in retaining the motivational nature of normative belief. On this view—where *akrasia* is defined with reference to partial belief, desire, and expected utility—it is clear that *akrasia* is not merely a breakdown of rationality. It is better understood as a problem of lack of confidence in one's normative beliefs. This account thereby also leaves us with the ability to *help* the *akratic* overcome their *akrasia* by helping them to raise their own credence in correct normative beliefs and lower their credence in incorrect ones. It should come as no surprise to those familiar with Ramsey and his version of pragmatism that his account of belief would leave us uniquely positioned to help the *akratic*, as his theories are structured around the position that "[p]hilosophy must be of some use to us [by] clear[ing] our thoughts and so our actions."[35] Aristotle, I think, would firmly agree with this principle, as he makes it very clear that idle contemplation of *akrasia* and virtue is pointless because we do not just want abstract knowledge. "We want to actually become good."[36]

---

[33] Ramsey, *Truth and Probability*, 80.

[34] Because *akrasia* is about normative propositions, not ones like 'that I exist,' pointing to such cases does not vindicate Aristotle's condition. Even if it were the case that we can be certain of propositions like the cogito, what would be needed is a normative proposition we are certain of.

[35] Ramsey, Frank. (1929b/1990) 'Philosophy'. Ramsey 1990, in Philosophical Papers. Ed. D. H. Mellor. Cambridge: Cambridge University Press, 1-8, 1.

[36] Aristotle, *Nicomachean Ethics*, 1103b25-30.

References

Aristotle (2020) *Nicomachean Ethics*. Adam Beresford, trans. London: Penguin Books.

Hume, David (1739) *A Treatise of Human Nature*. Oxford University Press.

MacBride, Fraser, Mathieu Marion, Frápolli, María José, Dorothy Edgington, Edward Elliott, Sebastian Lutz, and Jeffrey Paris (2019) 'Frank Ramsey'. https://plato.stanford.edu/entries/ramsey.

Misak, Cheryl (2018) *Cambridge Pragmatism*. Oxford University Press.

Peirce, Charles S (1877) 'The Fixation of Belief'. *Popular Science Monthly* **12**: 1-15.

Ramsey, Frank (1930/1991). *On Truth*. N Rescher and U Majer, eds. Kluwer. DOI: 10.1007/978-94-011-3738-6

Ramsey, Frank. (1927/1990) 'Facts and Propositions' in D H Mellor, ed., *Philosophical Papers*: 34-51. Cambridge University Press.

Ramsey, Frank. (1926/1990) 'Truth and Probability' in D H Mellor, ed., *Philosophical Papers*: 52-95. Cambridge University Press.

Ramsey, Frank. (1929a/1990) 'General Propositions and Causality' in D H Mellor, ed., *Philosophical Papers*: 145-64. Cambridge University Press.

Ramsey, Frank. (1929b/1990) 'Philosophy' in D H Mellor, ed., *Philosophical Papers*: 1-8. Cambridge University Press.