ISSN 1333-1108

Croatian Journal of Philosophy Vol. XXVI · No. 76 · 2026

Institute of Philosophy

Croatian Journal of Philosophy

Vol. XXVI · No. 76 · 2026

Croatian Journal of Philosophy

1333-1108 (Print)
1847-6139 (Online)

Published by
Institute of Philosophy
Ulica grada Vukovara 54/IV, 10000 Zagreb, Croatia
www.ifzg.hr

Available online at https://cjp.ifzg.hr
and https://hrcak.srce.hr/en/cjp

Croatian Journal of Philosophy is published three times a year. It publishes original scientific papers in the field of philosophy.

Croatian Journal of Philosophy is indexed in *The Philosopher's Index, PhilPapers, Scopus, ERIH PLUS* and in *Arts & Humanities Citation Index* (*Web of Science*).

Croatian Journal of Philosophy is published with the support of the Ministry of Science, Education and Youth of the Republic of Croatia.

Instructions for Contributors

All submissions should be made through our online submission system: https://ojs.srce.hr/index.php/cjp/about/submissions. Submissions must be in English and formatted to be double-spaced with suitably wide margins, an A4 page size, and automatic page numbering.

Articles are normally no longer than 8,000 words of main text including bibliography. The Journal will consider longer papers, but, once these limits are exceeded, authors should bear in mind the editorial policy that the acceptance bar raises with increasing length.

Manuscripts should be compiled in the following order: cover page; title; abstract (not exceeding 200 words); keywords (3 to 6); main text; appendices (as appropriate); references.

All the authors of a paper should include their full names, affiliations, postal addresses, telephone and fax numbers and email addresses on the cover page of the manuscript. If a paper is co-written, one author should be identified as the Corresponding Author. The cover page must be submitted as a separate document. All submitted manuscripts must be prepared for blind review, with revealing acknowledgements and self-identifying references removed.

Sources are cited in the text by the author's last name, the publication date of the work cited, and a page number if needed, e.g. (Barber 2007: 324). Full details appear in the reference list in which the year of publication appears immediately after the author's name:

Barber, A. 2007. "Linguistic Structure and the Brain." *Croatian Journal of Philosophy* 21 (7): 317–341.

Williamson, T. 2013. *Identity and Discrimination*. Oxford: Wiley-Blackwell.

The publication of a manuscript in the Croatian Journal of Philosophy is expected to follow standards of ethical behavior for all parties involved in the publishing process: authors, editors, and reviewers. The journal follows the principles of the Committee on Publication Ethics (https://publicationethics.org/resources/flowcharts).

CROATIAN JOURNAL OF PHILOSOPHY

Vol. XXVI · No. 76 · 2026

Introduction

The first issue of this year of the Croatian Journal of Philosophy *opens with a special section devoted to* The Behavioral and Ethical Consequences of Large Language Models. *This was also the theme of the* 3rd Kathy Wilkes Memorial Conference*, held in Turin, Italy, on 26–27 April 2024. I am grateful to the conference organisers, and in particular to Ms Nada Bruer of the Inter-University Centre Dubrovnik, for their assistance in preparing this section. The conference opened with welcome remarks by Monica Bucciarelli, Anita Avramides, Nada Bruer, and Riccardo Viale, and included presentations by Edoardo Datteri, Lukas J. Meier, Mario Rasetti, David Papineau, and Tvrtko Tadić. The present section brings together papers by Edoardo Datteri, Lukas J. Meier, and Riccardo Viale and Shaun Gallagher. In "Enactive Problem Solving and Chatbot Architectures," Riccardo Viale and Shaun Gallagher examine competing conceptions of rationality in decision-making and problem solving, and develop the notion of enactive problem solving (EPS). They argue that problem solving is best understood as an action-oriented and embodied process grounded in agent–environment interaction. Applied to therapeutic chatbots, this approach suggests that less rigid systems may, in certain respects, better support users' relational autonomy than more highly structured architectures. In "The Attribution of Rationality to Robots," Edoardo Datteri critically assesses current experimental approaches to the study of the intentional stance towards robots. He argues that the existing focus on the attribution of mental states neglects the attribution of rationality, thereby limiting our understanding of how users form expectations about robotic behaviour. In "Large Language Models versus Fuzzy Cognitive Maps for Solving Moral Dilemmas," Lukas J. Meier compares the performance of large language model–based systems with an alternative approach based on fuzzy cognitive maps (the METHAD algorithm) in the domain of medical ethics. He identifies systematic differences in interpretability, structure, and types of error, and suggests that an optimal advisory system would combine features of both approaches.*

Continuing from the special section, the issue proceeds with a set of articles addressing a range of topics in contemporary philosophy. Walter Veit and Heather Browning, in "Two Kinds of Conceptual Engineering," develop a pluralist account of conceptual engineering by distinguishing between its naturalist and moral variants. Rather than differing in method, these approaches are shown to diverge in their roles and purposes. Through the examples of health and animal welfare, the authors

illustrate how tensions between them can both generate conflict and contribute to conceptual improvement. A defence of libertarian free will is advanced by Davor Pećnjak in "Deliberation, Action and Freedom." Drawing on phenomenological considerations, Anselmian insights, and recent empirical findings concerning the inhibition of action, the paper presents a cumulative argument for the existence of genuine alternatives in both deliberation and action. Timothy J. Nulty's "Temporal Integration and the Basis of Moral Equality" offers a critical challenge to the widely accepted view that all persons possess equal moral status. On the basis of an account of temporal integration, Nulty argues that differences in individuals' connections to their future selves bear directly on morally relevant properties such as autonomy and agency, thereby complicating standard defences of moral equality. The issue concludes with Jinghua Chen's "Theoretical Sources of Rawls's Justice as Fairness: Kant, Hegel and Mill," which situates Rawls's theory within its broader philosophical context. The paper identifies important parallels between Rawls's central ideas—the original position, the primacy of the basic structure, and the two principles of justice—and corresponding elements in the thought of Kant, Hegel, and Mill.

TVRTKO JOLIĆ

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.1
Received: June 29, 2025
Accepted: October 15, 2025

Enactive Problem Solving and Chatbot Architectures

RICCARDO VIALE
Herbert Simon Society and University of Milano Bicocca, Milano, Italy

SHAUN GALLAGHER
University of Memphis, Memphis, USA

In this paper we review different conceptions of rationality as they apply to decision making and problem solving. We defend the notion of enactive problem solving (EPS) and we consider how it applies to the use of therapeutic chatbots. We start by recounting the development of different models of rationality, from expected value maximization, to subjective expected utility theory, and of bounded rationality from prospect theory, to simple adaptive heuristics. We then focus on problem solving. We show how the centre of gravity of problem solving shifts from the model of computational and cognitive processing, to an action-oriented pragmatic one in EPS, that is, to the possible actions that the body-environment interaction allows. We highlight a notion of the agent's relational autonomy as part of EPS and show that this can be a problem in some chatbot designs in the therapeutic context. EPS suggests that simple early versions of therapeutic chatbots like ELIZA may be more beneficial than more recent sophisticated models involving "Belief-Desire-Intention" (BDI) architecture characterized by predetermined trajectories. The system formed between the agent plus the chatbot tool facilitates a form of enactive problem solving, reducing rigidity on the side of the agent.

Keywords: Bounded rationality; enactive problem solving; chatbot; AI.

1. The repair programme about the limits of rationality

Questions about the limits of rationality are part of a long tradition that includes the 18th century critique of the theory of expected value maximization. The concept of expected value maximization, within the

context of decision theory, is credited to Blaise Pascal and Pierre de Fermat in the 17th century. They recognized that by calculating the expected value of different outcomes, a decision maker could choose the rational action that offered the highest potential payoff. Since then it has been a history of discovering empirical and logical anomalies in the theory. Gigerenzer (2008: 90), for example, has described the historical progression – from expected value maximization (as a standard of rationality) to expected utility theory and then on to prospect theory – as a "repair program" aimed at resuscitating the mathematical operation of weighted integration, based on the definition of mathematical expectation, as a psychological theory of rationality. Although expected value maximization was once regarded as a proper standard of rationality, it ran aground on the St. Petersburg Paradox, and Daniel Bernoulli began the repair program by transforming the outcomes associated with lotteries using a logarithmic utility of money function. This modification survived and grew as expected utility theory which took root in 20th century neoclassical economics. Then came Allais' Paradox, which damaged expected utility theory's ability to explain observed behavior, and a new repair appeared in the form of prospect theory, which introduced more transformations with additional parameters to square the basic operation of probability-weighted averaging with observed choices over lotteries.

2. Rationality as SEU decision making

The empirical study of human rationality from the post-war period to date has mainly developed by looking at a normative model of decision making. In particular Subjective Expected Utility (SEU) decision making, which stems from the subjective expected utility theory of Savage (1950) that itself extended the results of Von Neumann and Morgersten (1944).[1]

In decision theory, the von Neumann–Morgenstern utility theorem[2] shows that under certain axioms of rational behaviour, such as completeness and transitivity, a decision maker faced with risky (probabilistic) outcomes of different choices will behave as if he or she is maximizing the expected value of some function defined over the potential outcomes at some specified point in the future. The theory recommends which option rational individuals should choose in a complex situation, based on their risk appetite and preferences. The theory of subjective expected utility combines two concepts: first, a personal utility func-

[1] The way in which this escalation developed is discussed in detail in Mousavi and Tideman (2019).

[2] Von Neumann and Morgenstern never intended axiomatic rationality to describe what humans and other animals do or what they should do. Rather, their intention was to prove that if an individual satisfies the set of axioms, then their choice can be represented by a utility function.

tion, and second a personal probability distribution (usually based on Bayesian probability theory).[3]

In other words (Viale 2024), if decision makers want to reach a goal, for example enrolment at university, they face a limited number of alternatives (such as different universities and different courses), each of which has a probable outcome (such as cost, teaching value, difficulties of exams, type of professional training, consequences on the potential job i.e., on their salary, on the location of the employer, and so on) that has consequences related to the utility function. The concepts used to define the decision are therefore information about the world, the risk related to outcomes and consequences; preferences over alternatives; the relative utilities on the consequences; and, finally, the computation to maximize the subjective expected utility. Even if in formal decision theory no explicit reference is made to the actual empirical mental and psychological characteristics of the decision maker, in fact these can be mapped onto psychological processes such as external perceptual incoming inputs or internal mnemonic inputs, in mental representations of the states of the world on the basis of information, in hedonic evaluations[4] of the states of the world, and in deductive and probabilistic computation on the possible decisions to be implemented on the basis of hedonic evaluations.

3. SEU driven psychology of decision and bounded rationality

On this view the cognitive psychology of decision making precisely reflects the conceptual structure of formal decision theory. In relation to this structure and the normative component derived from it, empirical research in the cognitive psychology of decision making has been developing since the 1950s. Weiss and Shanteau (2021), highlight that in the 1950s Ward Edwards, the founder of the psychology of decision making, began to carry out laboratory experiments to unravel the way in which people actually decide. His experiments, which became the reference

[3] This theoretical model has been known for its clear and elegant structure and it's considered by some researchers one of "the most brilliant axiomatic theor[ies] of utility ever developed". Assuming the probability of an event, Savage defines it in terms of preferences over acts. Savage used the states (which are not under one's control) to calculate the probability of an event. On the other hand, he used utility and intrinsic preferences to predict the event's outcome. Savage assumed that each act and state are enough to uniquely determine an outcome. However, this assumption breaks down in cases where the individual lacks sufficient information about the event. In reality Savage explicitly limited the theory to small worlds, that is, situations in which the exhaustive and mutually exclusive set of future states S and their consequences C are known.

[4] The hedonic approach to economic assessment can be used for evaluating the economic value of goods. The hedonic approach is based on the assumption that goods can be considered aggregates of different attributes, some of which, as they cannot be sold separately, do not have an individual price.

of subsequent generations and in particular of Daniel Kahneman and Amos Tversky's Heuristics and Biases program, have two fundamental characteristics: firstly, the provisions of the SEU are set as a normative reference, and the experimental work has the aim of evaluating when and how the human decision maker deviates from the requirements of the SEU. Ultimately, the aim is to discover the irrational performances in the decision. Secondly, the experiments are not carried out in the real decision-making contexts of everyday life, but in an abstract one of games, gambling, bets and lotteries. In these abstract experimental situations, characterized by risk, the informative characteristics typical of the real environment – such as uncertainty, complexity, poor definition of data, instability of phenomena, dynamic and interactive change with the decision maker, and so on – are entirely absent. In fact, the gambling paradigm was considered the indispensable "fruit fly" of research on decision making (Lopes 1983) in those years, and experimental activity was confined to abstract situations characterized by risk, that is, by a defined and closed set of alternatives with outputs measurable at a probabilistic level and assessable at the level of relative utility. The influence of economics and its methodological imprint linked to physics is evident. Abstraction and simplification are the methodological objectives that allow independent variables to be kept under control and disturbance phenomena to be reduced (Blaug 1980; Viale 1997, 2012, 2013, forthcoming). However, it should not be overlooked that this may lead to the description of psychological abstractions that do not correspond to what happens in the individual's action processes in the reality of daily and professional life.

The empirical and deductive anomalies of the theory of rationality have informed the development of different concepts of bounded rationality. The limitations were found in the suboptimality of the information collection, in suboptimal computational ability to process data, and in the adaptive interaction between mind and environment. Robert Aumann (1962; 1997) advanced five arguments for bounded rationality that represent the different concepts. To summarize his remarks: (1) Even in very simple decision problems, most economic agents are not (deliberate) maximizers; (2) Even if economic agents aspired to pick a maximal element from a choice set, performing such maximizations are typically difficult and most people are unable to do so in practice; (3) Experiments indicate that people fail to satisfy the basic assumptions of rational decision theory; (4) Experiments indicate that the conclusions of rational analysis (broadly construed to include rational decision theory) do not match observed behavior; (5) Some of the conclusions of rational analysis do not agree with a reasonable normative standard.

Accordingly, for a majority of economic researchers across disciplines, bounded rationality is identified with some form of optimization problem under constraints. On the contrary, Gerd Gigerenzer is among the most prominent and vocal critics of the roles that optimization

methods and logical consistency play in commonplace normative standards for human rationality (Gigerenzer and Brighton 2009), especially the role those standards play in Kahneman and Tversky's biases and heuristics program (Kahneman and Tversky 1996; Gigerenzer 1996).

4. Problem solving action

Herbert Simon (1986) emphasizes the importance of problem solving and differentiates it from decision making, which he considers a phase downstream of the former. In dealing with a task, humans have to frame problems, set goals and develop alternatives. Evaluations and judgments about the future effects of the choice are the final stages of the cognitive activity. In fact, Simon's research in AI, economic and organizational theory is almost entirely dedicated to problem solving that seems to absorb the evaluation and judgment phase. On the one hand, action, to the extent that it has a role in this process, seems to arise at the end of the problem solving without interruption.

On the other hand, his approach to problem solving highlights the influence of American pragmatism, and in particular of John Dewey (1910) and Charles Sanders Peirce (1931) and William James (1890), on his work. The centre of gravity of the rationality of the action lies in the ability to adapt. And the centre of gravity of adaptation is not so much in the internal environment of the actor, that is, in his or her cognitive characteristics, as in the pragmatic external environment. Simon and Newell write: "For a system to be adaptive means that it is capable of grappling with whatever task environment confronts it. Hence, to the extent that a system is adaptive, its behaviour is determined by the demands of the task environment rather than by its own internal characteristics. Only when the environment stresses its capacities along some dimension – presses its performance to the limit – do we discover what those capabilities and limits are, and are we able to measure some of their parameters" (Simon and Newell 1971: 149).

The metaphor of the ant on the beach (Simon 1981) is illuminating: imagine an ant walking on a beach. Now let's say you wanted to understand why the ant is walking in the particular path that it is. In Simon's parable, you cannot understand the ant's behaviour just by looking at the ant: "Viewed as a geometric figure, the ant's path is irregular, complex, hard to describe. But its complexity is really a complexity in the surface of the beach, not a complexity in the ant." (Simon, 1981 (1988): 80). In other words, to predict the path of the ant, we have to consider the effects of the beach – the context that the ant is operating in. The message is clear: we cannot study what individuals want, need or value detached from the context of the environment that they are in. That environment shapes and influences their behaviour. In this example, the procedural rationality of the ant (finding a suitable behaviour on the beach) allows its substantial rationality (the adaptivity to the irregularity of the beach).

From this metaphor Simon derives a philosophical principle very much in tune with wide embodied cognition: "A man considered as a system capable of having a behaviour is very simple. The apparent complexity of his behaviour over time is largely a reflection of the complexity of the environment in which he finds himself" (Simon 1981 (1988): 81) The behaviour adapts to external purposes and reveals those characteristics of the system that limit its adaptation. Changes in the environment alter the ability to interact with it.

On this interpretation, the correspondence between action and solution of a problem conceptually bypasses the analytic phase of the decision and limits the role of symbolic representation (Viale 2024). The decision-making model based on SEU theory does not correspond to the empirical reality of individual action. In solving any problem, whether opening a door, running to catch a falling ball, replacing a car tyre, calculating for a financial investment, solving tests and puzzles or negotiating with a competitor, the search for the solution corresponds to acting, to a recursive feedback process leading up to the final action.

5. Enactive Problem Solving (EPS)

From the postwar period to today, the study of reasoning, judgment, and decision processes has largely been conducted within the classic cognitive model, often referred to as "Information Processing Psychology". It has three characteristics: first, thought is computation and occurs as computation. Every mental activity is performed by algorithms similar to the computer's machine language. Cognition derives from computational procedures that are carried out on abstract symbolic structures.

Cognitivism has portrayed the mind that thinks and decides as if it were in a vat, separated from the body and the environment. The mind is "disembodied" from the body that carries it and "detached" from the environment in which it interacts. The new perspective of 4E (embodied, embedded, extended, and enactive) cognition reveals instead a cognition integrated with the body through action and shaped by the environment with which the body interacts and in which it is located. This bodily interaction with the environment shapes and models the cognitive activity. Enactive approaches emphasize the idea that the body is dynamically coupled to the environment is important ways (Thompson 2007; Di Paolo 2005); they point not only to sensorimotor contingencies (where specific kinds of movement change perceptual input) (O'Regan and Noë 2001), but also to bodily affectivity and emotion (Colombetti 2013) as playing a nonrepresentational role in cognition. Embedded and enactive approaches emphasize action affordances that are body- and skill-relative (Chemero 2009). More generally, most theorists of embodied cognition hold that these ideas help to shift the ground away from orthodox, purely computational cognitive science, which clearly

informs the cognitive psychology of decision making. In this respect, it's not just the internal processes of the mind or brain, but the brain-body-environment system that is the unit of explanation.

For EPS action is part of the cognitive process (Viale, Gallagher, and Gallese 2023); deliberation, judgment and evaluation are intrinsic aspects of an action process that may involve the agent/problem solver moving around to get a better look, or actively engaging with others to gain perspective. Problem solving is a dynamic process based on pragmatic, recursive actions and on positive or negative feedback from the environment and its affordances. New possibilities, and new requirements emerge, not simply from the mind of the problem solver, but are often generated from the social interactions and worldly practices that characterize this situation. These are the factors that feed back into the problem-solving process. The enactive emphasis on flexibility in everyday problem solving emphasizes embodied and social practices – actual doings and material engagements that typically involve others in either direct or indirect ways – rather than internal deliberations, models, representations or rules. Deliberation is not ruled out, but it often occurs in consultation with others as problematic situations develop and as we work on a task, with the working component entering into deliberative practices. Models are often embodied in external drawings or diagrams or material arrangements that we try out as we try to solve the problem. Cognitive processes are extended by the tools and technologies we use. This conception of EPS delivers the action-oriented detail on the kind of adaptivity that Newell and Simon were seeking in the concept of bounded rationality – what they referred to as the "grappling with whatever task environment confronts" the agent (Newell and Simon 1971: 149), where 'grappling' means, for EPS, embodied hands-on activities that together with neural and extra-neural processes do the work of cognition.

One important principle central to the enactive view of problem solving is the agent's preservation of autonomy. Even when the agent encounters environmental constraints, problem solving continues only if the agent can improvise or make a move to resolve or make use of that constraint. Autonomy on this model is relational (Gallagher 2020; MacKenzie and Stoljar 2000) rather than implying absolute control. Nonetheless, some degree of autonomy is required for the process to continue.

The centre of gravity of problem solving is therefore no longer located in the computational and cognitive part, but it shifts to the action-oriented pragmatic one, that is, to the possible actions that the body-environment interaction allows (Viale 2024). This position, which places the constraints of the rational activity of choice and decision not so much in the computational possibilities of the human mind as in the mind-body-environment interaction, represents a further development of Herbert Simon's theory of Bounded Rationality (Viale, Gallagher,

and Gallese 2023). The environment cannot be analysed only as a structure of the task through its computational variables[5]. The physical and social environment also generates sensory and motor constraints that influence reasoning and action. And in determining a choice, possible or simulated bodily actions have an influence in shaping the range of possible options and the value attributed to them.

Relevant to the idea of problem solving, there is general agreement that the environment scaffolds our cognitive processes, and that our engagement with the environmental structure, and environmental features, including external props and devices, can shift cognitive load. Already, within the scope of Simon's own work it's clear that only through the enactive interaction between problem solver and environmental affordances is it possible to construct a solution.

For the idea of enactive problem solving, however, it is important to emphasize two things. First, the relational nature of affordances (Viale, Gallagher, and Gallese 2023). It is not just the environment that constrains behaviour; it is also the body's morphology and motor possibilities, and the agent's past experience and skill level that will define what counts as an affordance. The way in which the body couples (or can couple) to the environment, will delineate the set of possibilities or solutions available to the agent. Likewise, affordances can also be limited by an agent's affective processes, emotional states, and moods. It is sometimes not just what "I can" do (given my skill level and what the environment affords), but what "I feel like (or don't feel like)" doing (given my emotional state).

Second, as the pragmatists pointed out, the environment is not just the physical surroundings; it's also social and cultural and characterized by normative structures (Viale, Gallagher, and Gallese 2023). As Gibson (1977; 1979) indicated, affordances can be social. Enactive problem solving also highlights the important role of social and intersubjective interactions. Again, it's not only what "I can" do, but also what "I can't" (or "I ought not") do given normative or institutional constraints, as well as cultural factors that have to do with, for example, gender and race. These are larger issues that range from understanding how dyadic interactions shape our developing skills, to how institutional factors can either enable or constrain our social interactions.

6. BDI Artificial Intelligence

The first phase of AI can be described as the growth of so-called Simulative AI involving amongst other things descriptive protocols for com-

[5] For example, this includes the characteristics of the structure of the environment introduced by Gigerenzer and colleagues (Gigerenzer and Gassmaier 2011) such as uncertainty, redundancy, variability, number of alternatives and sample size. These characteristics derive from symbolically deconstructed empirical phenomena that are manipulated as cues with statistical meaning (tallied, weighted, sequenced and ordered).

puter simulation of human problem solving and thinking, based on the development of information processing psychology. It found its success in the development of expert systems aimed at problem solving and the ex-post simulation of scientific discoveries. It failed though in machine translation, image and speech processing and artificial movement.

One of the greatest successes of the first phase was the "logic theorist" and the general problem solver (GPS) computer program developed by Herbert Simon, Alan Newell and Cliff Shaw in 1956. The GPS was a computer program intended to work as a universal problem solver machine. In contrast to the former Logic Theorist project, the GPS worked with means–ends analysis. The first-generation AI programmes were psychology-based simulations of how humans reason and solve problems. They relied on a Cartesian dualism between mind and body (software and hardware), were fully transparent and predictable. They relied on a model of mind taken from folk psychology (Stich 1985) the belief–desire–intention triad (BDI). This model expressed by information processing psychology and 'good-old-fashioned' cognitivism is present in both SEU decision making and in computational models of problem solving.

The BDI model has been used to program intelligent agents. Superficially characterized by the implementation of an agent's beliefs, desires and intentions, it actually uses these concepts to solve a particular problem in agent programming. The BDI model allows for the representation of the characteristics and methods of achieving a goal in a system built according to the paradigms of software agents. Achieving the goal is what the agent works for. This is identified by a progression of its internal states that tends to implement the task in a stable manner by determining the actions that the agent is able to undertake. The simplest implementation of this concept is an algorithm that returns a value indicating success if the goal is achieved. However, more complex goals can be divided into different categories:

- achievement goals (long-term goals)
- satisfaction goals (recurring goals, such as gathering resources)
- preservation goals (preservation of life and property)
- delta goals (for example, changes in state).

In the BDI approach, intentions and motivational aspects are also included in the achievement of the goal.

These architectures are based on the concept of practical reasoning (Bratman 1999) which allows the analysis of the reasoning process performed by people when they aim to satisfy their expectations in the real world. The aim is to create an artificial system that simulates such behavior. A fundamental moment of practical reasoning is the choice of one's own objectives among those that are believed to be achievable. This choice process presupposes knowledge of both the options available and one's own desires, aspects involved in the traditional concept of autonomy (Frankfurt 1988). The decisions taken have different de-

grees of complexity and importance: fundamental choices can be made, such as deciding to exploit a totally new and unexpected opportunity, or simply deciding the best way to do something. In the theory of practical reasoning this step is described as the transformation of the chosen options into intentions, which have the characteristic of persisting as long as our beliefs make them reasonable. Practical reasoning is composed of two main processes:

- the decision process between a set of different perspectives
- the process for achieving the proposed condition.

In the process described, proactive behavior (persistence of intentions) and reactive behavior (abandonment of intentions because they are no longer convenient and adoption of a new goal) can be clearly identified. Reactive behavior allows us to verify our choices in light of the latest information, instead of blindly focusing on what we have previously decided. The fundamental characteristic of BDI architectures is to use data structures that correspond to the beliefs, desires and intentions referred to in the theory of practical reasoning; decision-making processes are implemented by functions that act on these structures.

7. Chatbot architecture: BDI versus EPS

We consider here the use of chatbots in the therapeutic context, motivated in some cases by problems with intersubjective relations where patients mistrust the therapist and resist therapy, leading to high dropout rates (Martino et al. 2012). AI-supported interventions (in the form of a chatbot) are proposed to address these issues (Szalai 2021). The specific proposal: AI programs in natural language processing "could help patients re-author their self-narratives into more coherent and meaningful sequences, in which they view themselves more like agents and less subject to external control, as well as assign more consistent roles to others" (Szalai 2021). The aim, in part, is to enhance the autonomy of the patient.

The chatbot is a natural language processing-based AI device that is designed to mimic human interaction in order to supplement sessions with the therapist. The system is meant to facilitate the patient's reflective processes leading to the reframing of self-narrative. The benefit of such a system is that it could reduce motivation for the patient's antagonistic responses to the therapist. Szalai notes that patients are more willing to reveal information to "virtual humans" than to the human therapist. The use of AI in this context is not meant to solve all problems; nor is it meant to replace the therapist, and it cannot be the complete therapeutic process.

One of the early first attempts at creating a natural language program that could run a simple psychotherapeutic encounter was ELIZA, a rule-based chat system that would basically turn anything you said into a question that was repeated back to you, with the idea of getting

you to elucidate. If all goes well, you seemingly enact or talk your way to a cure. ELIZA doesn't diagnose anything. Rather it simply emulates a Carl Rogers style of talk therapy that mirrors whatever the analysand says (Bassett 2019). It runs a rule that basically asks you to elucidate, with the goal of "self-actualization." Although ELIZA seems somewhat primitive compared to current AI deep-learning chatbots, there are anecdotal reports that subjects appreciated the encounter. As Caroline Bassett puts it, "if humans found ELIZA useful perhaps it was as a mirror, a listening surface which enabled forms of self-examination, self expression, or self re-narrativization. If users found something revealing in their interactions with ELIZA then that something was their own: ELIZA never did, and does not now, deliver injunctions, suggestions—or nudges; and has no program to promulgate" (2019: 809). ELIZA itself doesn't have a goal, and there is no trajectory involved; it's not committed to anything. As such, it grants a high degree of autonomy to the person using it. We suggest that this is a model that is consistent with enactive problem solving. The therapeutic solution emerges from the patient's interaction with the technology in a way that preserves the patient's autonomy.

In contrast to ELIZA, current chatbots can be designed with a goal or trajectory built in. For example, a BDI architecture tends to rigidly stay on course, or "overcommit" to a predetermined goal (Wallis 2022) – this kind of inflexibility could lead patients to an impersonally predetermined self-construction – a situation that can rob the patient of autonomy.[6] The rigidity noted by Wallis is not easily fixed and may be persistent in such systems, introducing issues of sustainability. Artificial systems may fail or may work differently than expected in changing circumstances (so-called counterperformativity (Bamford and MacKenzie 2018)). Circumstances (including human inclinations) frequently change.

An alternative approach, proposed by Wallis, drawing on enactivist ideas, is to design a system to directly recognize a trajectory (a goal or intention) in the patient's action by recognizing when its own behavior could be complementary, thus forming a shared trajectory. The task is to get the AI system or artificial agent to recognize what the patient's action or expression affords, i.e., to register it in terms of its own possible response with a high probability weighting for appropriateness. Wallis admits, however, that "implementing this on a computer is beyond us at the moment…" (2022: 9). This, however, is not just a pragmatic issue – perhaps to be overcome in the fast pace of AI

[6] Chatbots offer a form of cognitive scaffolding as a structure meant to support or enable the user's problem solving. One important consideration here is whether these trajectories are transparent to the user. If the user is unaware of how the chatbot is designed for a particular outcome, and how it may be shaping the user's thoughts or experiences, this threatens the user's autonomy. This is sometimes referred to as 'hostile scaffolding' (Timms and Spurrett 2023)

progress – or a theoretical problem that requires more research – it's a tricky kind of hermeneutical problem. If one allows the AI system to be guided by the patient, the shared trajectory may simply become a reinforced loop back into the patient's disorder. On this score, however, the same worry applies to ELIZA. This, we note, is a common worry in the therapeutic context – how to allow some significant degree of autonomy (or agency) to the patient, but at the same time effect a solution that moves the patient to a new perspective. In this respect, the solution may depend on the intervention (a meta-guidance) by a seemingly irreplaceable human psychotherapist who is attuned to counter problems caused by structural underdetermination (as in ELIZA) or overdetermination (as in BDI).

References

Aumann, R. J. 1962. "Utility Theory without the Completeness Axiom." *Econometrica* 30 (3): 445–462.

Aumann, R. J. 1997. "Rationality and Bounded Rationality." *Games and Economic Behavior* 21 (1–2): 2–17.

Bamford, A. and D. MacKenzie. 2018. "Counterperformativity." *New Left Review* 113: 97–121.

Bassett, C. 2019. "The Computational Therapeutic: Exploring Weizenbaum's ELIZA as a History of the Present." *AI & Society* 34 (4): 803–812.

Blaug, M. 1980. *The Methodology of Economics.* Cambridge: Cambridge University Press.

Bratman, M. E. 1999. *Intention, Plans, and Practical Reason.* Stanford: CSLI Publications.

Chemero, A. 2009. *Radical Embodied Cognitive Science.* Cambridge: MIT Press.

Colombetti, G. 2013. *The Feeling Body: Affective Science Meets the Enactive Mind.* Cambridge: MIT Press.

Dewey, J. 1910. *How We Think.* Boston: D. C. Heath.

Di Paolo, E. A. 2005. "Autopoiesis, Adaptivity, Teleology, Agency." *Phenomenology and the Cognitive Sciences* 4 (4): 429–452.

Frankfurt, H. G. 1988. *The Importance of What We Care About.* Cambridge: Cambridge University Press.

Gallagher, S. 2020. *Action and Interaction.* Oxford: Oxford University Press.

Gibson, J. J. 1977. "The Theory of Affordances." In R. Shaw and J. Bransford (eds.), *Perceiving, Acting, and Knowing.* Hillsdale: Erlbaum, 67–82.

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Gigerenzer, G. 1996. "On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky." *Psychological Review* 103 (3): 592–596.

Gigerenzer, G. 2008. *Rationality for Mortals: How People Cope with Uncertainty.* Oxford: Oxford University Press.

Gigerenzer, G. and H. Brighton. 2009. "Homo Heuristicus: Why Biased Minds Make Better Inferences." *Topics in Cognitive Science* 1 (1): 107–143.

Gigerenzer, G. and W. Gaissmaier. 2011. "Heuristic Decision Making." *Annual Review of Psychology* 62: 451–482.

Kahneman, D. and A. Tversky. 1996. "On the Reality of Cognitive Illusions." *Psychological Review* 103 (3): 582–591.

Lopes, L. L. 1983. "Some Thoughts on the Psychological Concept of Risk." *Journal of Experimental Psychology: Human Perception and Performance* 9: 137–144.

James, W. 1890. *The Principles of Psychology*. New York: Dover.

Mackenzie, C. and N. Stoljar (eds.). 2000. *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford: Oxford University Press.

Martino, F., M. Menchetti, E. Pozzi and D. Berardi. 2012. "Predictors of Dropout among Personality Disorders in a Specialist Outpatients Psychosocial Treatment." *Psychiatry and Clinical Neurosciences* 66 (3): 180–186.

Mousavi, S. and N. Tideman. 2021. "Beyond Economists' Armchair: The Rise of Procedural Economics." In R. Viale (ed.), *Routledge Handbook of Bounded Rationality*. London: Routledge.

Peirce, C. S. 1931. *Collected Papers of Charles Sanders Peirce*. Cambridge: Harvard University Press.

Savage, L. J. 1954. *The Foundations of Statistics*. 2nd ed. New York: Dover.

Simon, H. A. 1981. *The Sciences of the Artificial*. Cambridge: MIT Press.

Simon, H. A. 1986. *Decision Making and Problem Solving*. Washington, DC: National Academy Press.

Simon, H. A. and A. Newell. 1971. "Human Problem Solving: The State of the Theory in 1970." *American Psychologist* 26 (2): 145–159.

Stich, S. 1985. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, MA: MIT Press.

Szalai, J. 2021. "The Potential Use of Artificial Intelligence in the Therapy of Borderline Personality Disorder." *Journal of Evaluation in Clinical Practice* 27 (3): 491–496.

Thompson, E. 2007. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA: Harvard University Press.

Timms, R. and D. Spurrett. 2023. "Hostile Scaffolding." *Philosophical Papers* 52 (1): 53–82.

Von Neumann, J. and O. Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Viale, R. 1997. *Cognitive Economics*. Milan: Università Bocconi.

Viale, R. 2012. *Methodological Cognitivism: Mind, Rationality and Society*. Heidelberg: Springer.

Viale, R. 2013. *Methodological Cognitivism: Cognition, Science and Innovation*. Heidelberg: Springer.

Viale, R. 2024. "Enactive Problem Solving: An Alternative to Decision Making." In G. Gigerenzer et al. (eds.), *Companion to Herbert Simon*. Cheltenham: Elgar.

Viale, R. Forthcoming. "What Is Cognitive Economics?" In R. Viale (ed.). *Handbook of Cognitive Economics*. Cheltenham: Elgar.

Viale, R., S. Gallagher and V. Gallese. 2023. "Bounded Rationality, Enactive Problem Solving, and the Neuroscience of Social Interaction." *Frontiers in Psychology* 14: 1152866.

Wallis, P. 2022. "An Enactivist Account of Mind Reading in Natural Language Understanding." *Multimodal Technologies and Interaction* 6 (5): 32.

Weiss, D. J. and J. Shanteau. 2021. "The Futility of Decision Making Research." *Studies in History and Philosophy of Science* 90: 10–14.

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.2
Received: July 18, 2025
Accepted: February 15, 2026

The Attribution of Rationality to Robots

EDOARDO DATTERI
University of Milano-Bicocca, Milan, Italy

An increasing number of studies are attempting to determine, through quantitative experimentation, whether people adopt an intentional stance towards robots. These studies mainly use questionnaires in which participants are asked to choose between mentalistic and non-mentalistic descriptions of robotic behaviours portrayed in pictures. While these methods are extremely interesting in their attempt to operationalise Dennett's theoretical constructs, they only capture one aspect of the intentional stance: the attribution of mental states to robots. They neglect the question of whether participants also attribute rationality to the system. Consequently, they are not well equipped to analyse how people form expectations about the behaviour of the robots they interact with, which is crucial for studying the dynamics of human–robot interaction. There is indeed no reason to deny that laypeople might occasionally attribute mental states to robots while believing that they can act irrationally or model the decision-making processes of the system in terms devoid of any reference to rationality. Building on these considerations, this article reflects on an emerging area of research in human-robot interaction from a philosophical perspective, identifying a potential limitation that could be overcome by referring to psychological literature on the attribution of rationality to humans.

Keywords: Human-robot interaction; mental state attribution; intentional stance.

1. Introduction

When people interact with autonomous artificial systems such as intelligent chatbots and social robots, they sometimes *talk* about them in mentalistic terms ("Hey, look! The robotic vacuum cleaner *wants* to go to the kitchen!"). They can also *attribute* beliefs and desires to the system, assuming that it will behave rationally based on them. In a

series of groundbreaking essays, the philosopher Daniel Dennett famously termed this latter phenomenon the adoption of an 'intentional stance' towards the observed system, discussing various philosophical aspects of the idea. For example, he considered what role beliefs and desires play in intentional systems theory and whether the intentional stance aligns with common-sense psychology (Dennett 1989). The notion of intentional stance pervades contemporary research on human–robot interaction. In particular, a growing community of researchers comprising roboticists, psychologists and social scientists is studying the adoption of an intentional stance towards robots using a variety of experimental techniques (see Thellman et al. 2022 for a review). This article aims to comment on a particular aspect of this scientific enterprise by suggesting that this research focuses too much on attributing beliefs and desires to robots and neglects another important aspect of the intentional stance: attributing *rationality* to robots.

The concept of rationality plays a central role in Dennett's theory of intentional systems. He presents the intentional stance as involving the attribution of mental states, such as beliefs and desires, *as well as* rationality, to the target system. Adopting the intentional stance is not just about attributing the desire to reach the kitchen and the belief that the battery needs to recharge to a robotic vacuum cleaner, for example; it also involves assuming that there is a certain relationship between the system's beliefs and desires, and its behaviour – namely, that the system will act rationally based on them. However, it has been suggested that this assumption is too strong, as we sometimes observe irrational behaviour in systems to which we attribute a mind. For instance, people sometimes appear to fail to make the most rational choice in tasks such as the Wason selection task (Wason 1968). Consequently, there has been a debate about whether, when interpreting others' behaviour in mentalistic terms, people are actually adopting the intentional stance or engaging in a form of folk psychology that does not necessarily involve the attribution of rationality (see Dennett 1989; Stich 1981). This paper does not seek to take a position in this debate. It only argues that attributing mental states to a system is independent of attributing rationality to it. This means that the attribution of mental states can happen without attributing rationality, and vice versa. It also argues that current attempts to determine whether people adopt an intentional stance towards robots tend to focus only on attributing mental states. This means they neglect attributing rationality and do not actually investigate the adoption of an intentional stance towards robots. In doing so, this article provides a philosophical reflection on certain aspects of contemporary research in the fields of social robotics and human–robot interaction.

Why is it important to study whether people attribute not only a mind, but also rationality, to robots? Our understanding of these systems clearly influences our behaviour in their presence and how we interact with them. However, while certain aspects of our behaviour can

be affected by the mere attribution of mental states to robots, others arguably depend more deeply on whether we also attribute rationality to them. Therefore, to understand certain aspects of our interactions with robots, it is beneficial, if not essential, to determine whether we regard them as rational, irrational or non-rational. One study to be discussed in Section 3, by Wiese and colleagues (2012) suggests that adopting an intentional stance towards a robot activates certain cognitive processes, such as gaze following, which are not activated when that stance is not adopted. According to the authors, this occurs in a reflexive, bottom-up way that does not necessarily involve rationalising the robot's behaviour. Studying our attribution of rationality to robots may be less important for understanding our low-level reactions to their behaviour. However, other aspects of human–robot interaction depend more crucially on our expectations of their behaviour. For instance, when the 'battery' LED on our robotic vacuum cleaner lights up and the robot heads towards the kitchen, we anticipate that it will go to the docking station, which is indeed located there. If we then recall putting a bin in front of the docking station, we will move it to allow the robot to recharge properly. Similarly, if we are crossing a street and an autonomous car is approaching (Ziemke 2020), we will decide whether to continue or step back depending on whether we expect the car to stop at the crossing. Our decision to install an assistive robot in our elderly aunt's apartment will be based on our expectations of how the robot will behave, and whether or not it will harm her.

The issue is that the dynamics of our interactions with robots are significantly shaped by our expectations of their behaviour. In order to understand people's high-level reactions to robots and how they decide to behave when interacting with them, it is important to study how people predict robot behaviour.[1] In order to study people's predictions, it is important to determine not only whether they attribute a mind or specific mental states to robots, but also whether they assume that their behaviour will be rationally influenced by their beliefs and desires. In order to understand how pedestrians interact with autonomous cars, it is important to determine whether they attribute the desire not to hurt them to those cars; however, it is also important to determine whether they assume the car will act rationally based on that desire. The most rational thing for a car with this desire would be to stop at the crossing. However, the pedestrian's mental model of the car may include the assumption that it can act *irrationally* in certain situations, possibly because rationality is considered to apply only to

[1] Thellman and Ziemke (2020) have argued that more research in this area is needed. "Despite a growing interest in the role of mental state attribution in people's mental models of robots, and the importance of perceptual belief tracking in the context of social interaction, no research has so far targeted people's ability to predict the behavior of robots based on assumptions about how they perceive the environment." Their study is one of the few that explicitly addresses this important issue.

human beings. Alternatively, the pedestrian may simply assume that the car's behaviour is lawfully connected to its beliefs and desires in a way that is not characterised in terms of rationality or irrationality. People's mental models of robots may differ greatly in the way they connect attributed beliefs and desires to overt behaviour.

Various methods have been developed to study people's mental models of robots (Rueben et al. 2021), particularly in terms of whether people adopt the intentional stance towards them. Some studies have examined how this adoption affects brain activity (Chaminade et al. 2012), gaze following (Wiese et al. 2012), gaze aversion (Desideri et al. 2021), and other cognitive processes (Ciardo et al. 2020; Marchesi et al. 2025; Roselli et al. 2022). For reviews and general reflections, see Chaminade and Cheng (2009) and Wykowska and colleagues (2016). Other studies have attempted to determine whether people adopt the intentional stance based on the robot's physical appearance, behaviour, or other contextual factors (Mandell et al. 2017; Marchesi et al. 2019; Martini et al. 2015, 2016; Terada et al. 2007). These studies have made significant contributions to our understanding of how people perceive robots. However, they do not establish whether participants fully adopt the intentional stance towards robots since they only investigate the attribution of mental states to systems, not rationality. To demonstrate this, it is helpful to provide a brief overview of the intentional stance and argue that attributing beliefs and desires does not necessarily imply attributing rationality: a robot can be perceived as having beliefs and desires (and other propositional attitudes) while acting irrationally, or non-rationally, with respect to them.

2. Rationality and the intentional stance

2.1 The intentional stance

Dennett's theory of intentional systems is well known and has been the subject of extensive discussion since the publication of his seminal article (Dennett 1971). This section aims to recap some of its key features in preparation for the subsequent discussion. In this oft-quoted passage from (Dennett 1989), he defines the intentional stance (IS) as follows:

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many – but not all – instances yield a decision about what the agent ought to do; that is what you predict the agent *will* do. (17)

The IS is often presented by the author as an explanatory strategy (e.g. Dennett 1971, 2009). Suppose the 'battery' LED on a robotic vacuum cleaner lights up and the robot turns towards the kitchen. An observer

adopting the IS might explain this behaviour by attributing the following beliefs and desires to the robot: the belief that the battery is running out of charge; the belief that the docking station is in the kitchen; and the desire to recharge. Given these beliefs and desires, they would assume that the most rational thing for the robot to do would be to turn towards the kitchen. In this way, the IS provides a rational explanation for the robot's behaviour. The IS can also be regarded as a predictive strategy: the observer might predict that the most rational course of action, after turning right, would be to proceed straight towards the kitchen. However, it is important to stress that, in Dennett's framework, attributing beliefs and desires to the system in the IS does not mean believing that the system actually possesses them. The observer only treats the system as if it possesses them for the sole purpose of explaining and predicting its behaviour. Intentional systems theory is based on an instrumentalist conception of beliefs, desires, and other propositional attitudes. In this theory, the conditions under which the target system can be said to have a belief (or desire) with a particular content do not have to be the same as those adopted in other psychological frameworks based on the attribution of so-called propositional attitudes.

Rationality plays a central role in Dennett's theory of intentional systems. Firstly, the rationality of the system forms an integral part of the mental model constructed by the observer. It characterises the relationship between the system's beliefs and desires, and its behaviour (meaning that, according to the IS, the system will act rationally relative to its beliefs and desires). Secondly, rationality arguably influences the process by which the observer identifies the beliefs and desires that explain the system's behaviour. Why did the robot turn right? One first assumes the rationality of the system – in this case, that turning right was the most rational response to the situation. Then, one assumes that the system has beliefs and goals that make the right-turning behaviour rational, i.e. one makes "adjustments in the information-processing conditions" (Dennett 1971: 94). In this case, the aforementioned beliefs and desires (the belief that the battery is low and that the docking station is in the kitchen, and the desire to recharge) do indeed render this behaviour rational. According to intentional systems theory, explanation is akin to rationalisation: the assumption that the system is rational informs the observer's decision regarding the beliefs and desires attributed to the system.

Note that predictive tasks differ from explanations in that the behaviour of the system is unknown. To predict how the robot will act after turning right, the observer must first attribute desires and beliefs to the system and then assume that the resulting action is the most rational given these premises. However, as the action is still unknown, it is not possible to determine the system's beliefs and desires based on the assumption that the behaviour produced was rational, as in an explanatory context. According to Dennett (see the chapter "Three

Kinds of Intentional Psychology" in Dennett 1989), one ascribes the beliefs that the system ought to have "given its perceptual capacities, its epistemic needs, and its biography", as well as the desires that the system ought to have "given its biological needs and the most practicable means of satisfying them". Arguably, an assumption of rationality plays a role here too, albeit perhaps more covertly, because the system is assumed to have beliefs and desires that are rational, given its epistemic and biological needs.

In summary, when adopting an IS towards a system such as the robotic vacuum cleaner used as an example in this section, both mentalistic and rationality assumptions are made. The mentalistic assumption is that the system has beliefs, desires, and other mental states with content (conceived as "idealised fictions in an action-predicting, action-explaining calculus" Dennett 1978: 30), while the rationality assumption is that the system acts rationally based on them. It is through this rationality assumption that the observer can connect the system's beliefs and desires to its actual behaviour; the rationality assumption plays a key role in generating expectations about the robot's future behaviour.

It is worth noting that Dennett does not provide a clear definition of rationality. Stich (1981) notes that Dennett himself "has no illusions on the point" and "portrays intentional-systems theory – the general normative theory of rationality – as a discipline in its infancy" (42). Indeed, Dennett clearly acknowledges this in his work (Dennett 1989): "What then do I say rationality is? I don't say" (94). He also claims to have "good reasons for cautiously resisting the demand for a declaration on the nature of rationality" (94) and merely indicates what, in his view, rationality is not: it is neither deductive closure nor logical consistency. He never connects the intentional systems theory to classical decision theory, which involves ordering the set of possible actions by assigning probabilities to their potential outcomes and determining their expected utility in relation to the system's goals (Mele and Rawling 2004; Peterson 2017). Even if he did, there would be many possible ways to implement a classical decision-making process. Unless additional details were specified, including the definition of the set of possible actions and the criteria for calculating their consequences and assigning their expected utilities with respect to the goals, the attribution of rationality that accompanies the IS would not enable the observer to make precise predictions about the system's behaviour. This suggests that Dennett's presentation of the IS conceptual framework provides poor grounds for the analysis of people's behavioural expectations about robots unless some aspects of it are more precisely worked out.

2.2 Non-mentalistic rationalization

The next section will argue that, although tasks for investigating the attribution of rationality to systems have been developed in non-ro-

botic psychological literature (see Gergely et al. 1995), contemporary research on adopting an IS towards robots focuses narrowly on attributing mental states to them and neglects the phenomenon of rationality attribution. In support of this claim, it is worth noting that these two types of attribution are relatively independent of one another; one can rationalise a robot's behaviour without ascribing mental states to it, and vice versa. Without this further assumption, one might dispute the main assertion of this article, arguing that methods for studying the attribution of mental states to robots also determine whether participants attribute rationality to them.

Can a mental model of a robot incorporate the assumption that the robot is rational without also assuming that it has mental states? This depends on how the system is regarded as rational. According to the IS, the system acts rationally *with respect to its beliefs and desires*. Thus, in Dennett's framework, rationality cannot be attributed without also attributing mental states. However, a robot may be considered rational in ways that differ from the IS. For instance, an observer might assume that the robot adheres to a specific standard of rationality based solely on its behaviour: this concept is known as *rational analysis* (Anderson 1991). As Anderson presents it, rational analysis belongs to a long tradition of trying to understand behaviour as an adaptation to the system's environment.

> A rational analysis is an explanation of an aspect of human behavior based on the assumption that it is optimized somehow to the structure of the environment. […] As in economics, the term does not imply any actual logical deduction in choosing the optimal behavior, only that the behavior is optimized. (471)

Anderson outlines the various steps of rational analysis as follows: First, the goals of the system are specified. As any behaviour can be viewed as optimising a potential goal, external constraints must be considered in order to select the system's actual goal. Note that the system is assumed to have *goals*, not *desires*. This is evident from the rational analysis of the various systems that Anderson uses as examples. For instance, the goal of memory systems is to access necessary information from the past, while categorisation systems aim to predict features of objects. Anderson does not characterise these systems as having desires, conceived as intentional mental states. According to this approach, it is appropriate to say that the goal of a thermostat is to maintain the environmental temperature within a certain range, without implying that the thermostat desires this. The second step is to develop a formal model of the environment to which the system is adapted. Anderson specifies that this must be done from the system's perspective; the model must include environmental features that are accessible to the system rather than taking the observer's point of view as a reference. The third step is to make some minimal assumptions about the computational limitations of the system under analysis. For example, we might assume that the system's ability to process alterna-

tives is limited, that processing incurs a cost, or that its short-term memory is finite. The fourth step is to derive the system's optimal behaviour given these assumptions. This involves predicting behaviour that will maximise expected utility when considering the goals identified in step one, the environmental constraints defined in step two and the computational costs defined in step three. These predictions are then evaluated against the existing literature or experimental results. If necessary, the theory is revised iteratively.

Anderson points out that identifying the variables that the system is optimising may constrain the identification of the internal mechanisms responsible for its behaviour. Some authors in the Commentary on (Anderson 1991) have claimed that identifying the computational limitations to which the system is subject at step three requires one to consider possible internal mechanisms. Anderson's approach to rational analysis occasionally suggests that it can inform the discovery of mechanisms. For example, he states that his rational analysis of problem solving "seems more like the actual problem solving that people face daily" (481), and he outlines a procedure for selecting the partial plan with the greatest probability of success. However, he also states that "the structure driving explanation in a rational theory is that of the environment" rather than the mind of the system (471). The goal of rational analysis is "to predict behavior from the structure of the environment rather than the structure of the mind" (474) and "it is in the spirit of a rational analysis to prescribe what the behavior of a system should be rather than how to compute it". The question of how the system succeeds in producing optimal behaviour – whether through the interaction of beliefs, desires, intentions and other propositional attitudes, or by virtue of a different mechanism – is not one that rational analysis is intended to solve. The rational analysis of problem solving presented in his article is objective in the sense that it is not necessarily "in the subject's head" (482). Furthermore, he does not claim "that the human system actually goes through the relatively complex Bayesian analysis used to establish what the optimal behavior was" (483). The term 'rational' "is used in the economist's sense, which is that the output of the system is optimal and no claim is made about the mental processes by which this output is computed" (510).

Clearly, rational analysis can be applied to the behaviour of robots. In this case, the observer does not make any assumptions about the mental state of the system, even instrumentally. The observer only needs to make assumptions corresponding to the steps of rational analysis, as previously illustrated, and these steps do not include anything concerning the system's mental state. Rational analysis can particularly be combined with adopting a design stance towards the system based on the assumption that its functional and computational processes are adapted to its intended purpose. Although Anderson's rational analysis would not lead to the adoption of a rationality assumption as intended by Dennett (since the latter refers to the system's beliefs and desires),

the considerations made in this section suggest that it is, in principle, possible to consider a robot to be rational and to use this assumption to predict its future behaviour without necessarily assuming that it possesses any mental states.

2.3 Non-rational mentalization

Can mental states be attributed to a robot without also assuming that it always acts in the most rational way? There is no obvious reason why an observer should be prevented from assuming that the robot's decision-making system is occasionally or always irrational. Moreover, an alternative assumption could be made: that the behaviour of the system is neither rational nor irrational, but simply *non-rational*. This would mean that the relationship between mental states and system behaviour is modelled without any reference to rationality. For example, suppose the observer attributes the following beliefs and desires to the robotic vacuum cleaner: the belief that its battery is running low and that the docking station is in the kitchen, and the desire to recharge. As previously mentioned, these assumptions are idle for prediction purposes in themselves, unless the observer makes additional assumptions about how these mental states contribute to behaviour. The rationality assumption incorporated in the IS can be expressed as follows:

> (A) The system performs the most rational action with respect to its desires and beliefs.

As previously mentioned, it is unclear how this assumption can yield precise predictions of the robot's behaviour unless accompanied by a detailed account of what constitutes rational behaviour. As such, it offers only a generic representation of how the system's behaviour is influenced by its desires and beliefs. For the purposes of this discussion, the important point is that ascribing those beliefs and desires to the system does not imply that the observer must assume (A) or make equivalent attributions of rationality. It is not inconsistent to ascribe certain beliefs and desires to the system while assuming that it will behave irrationally relative to them. Indeed, the observer might well assume (instead of A) something along the following lines:

> (B) The system performs the least rational action with respect to its desires and beliefs.

Now, it is one thing to assume that the system acts *irrationally* and another to assume that it acts *non-rationally*. The latter occurs when the observer does not characterise the relationship between mental states and behaviour in terms of rationality or irrationality, but assumes that the robot's mental states are connected to its behaviour via law-like generalisations, for example:

> (C) If the system desires to recharge and believes that the docking station is in the kitchen, it will turn towards the kitchen.

Or, more generally:

(D) If the system desires to recharge and believes that the docking station is in room X, then it will turn towards room X.

The striking difference between option (A) and options (C) and (D) is that the latter ones do not explicitly refer to rationality. Here, the system's beliefs, desires and other propositional attitudes interact in a law-like manner to produce behaviour, with no assumption as to whether these regularities are rational or irrational. It should be noted that the point is not that these generalisations *cannot*, in principle, be said to be rational or irrational. In fact, turning towards the kitchen is intuitively rational with respect to the desire to recharge, and options (C) and (D) may be said to guide the system towards rational behaviour. However, the notion of rationality plays no essential role in predicting the behaviour of the system based on these two generalisations. To illustrate this, suppose that one of the law-like regularities attributed by the observer to the system has the following form:

(E) If the system has a belief with content *a* and a belief with content (if *a* then *b*), then the system forms a belief with content *b*.

This generalisation can be used to predict the presence of specific beliefs within a system's knowledge base. For instance, if an observer attributes two beliefs to the system – one stating that today is Friday, and the other stating that, if today is Friday, it is laundry day – the above regularity would enable the observer to predict that the system would possess the belief that today is laundry day. Other generalisations attributed to the robot could allow the observer to link the system's possession of this belief to its subsequent motor actions (e.g. doing the laundry). One could argue that this generalisation takes the form of a *modus ponens* inference rule, which would make the system's decision-making process appear rational according to a traditional view of rationality. In other words, one could assume that the system is regulated by (E) *and* that (E) is a rational decision-making rule. However, this additional assumption does not play a crucial predictive role. An observer unaware that (E) incorporates a principle of rational thinking, or who conceives of rationality differently, would still predict that the robot believes 'Today is laundry day'. Assumption (E) would support this prediction even if *modus ponens* were the hallmark of irrationality. In this case, the notion of rationality plays no role in the observer's mental model of the system. This non-rational approach to understanding the robot's behaviour could result in the following generalisation being attributed to it:

(F) If the system has beliefs with contents *b* and (if *a*, then *b*), then it forms a belief with the content *a*.

The fact that (F) is a fallacious rule of inference would likely prevent it from being included in IS-style rationalisations of the system's behaviour. However, nothing actually prevents one from modelling the inter-

action between the system's beliefs, desires and behaviour in terms of this law-like regularity, without any reference to rationality.

In terms of prediction, how different are these two strategies? While they may converge in some cases, in others they will diverge significantly. Suppose two observers both attribute the beliefs and desires mentioned before to a robotic vacuum cleaner. Observer 1 adopts an IS approach, assuming that the robot will act in the most rational way given its current beliefs and desires. Observer 2 assumes that the robot is subject to (D) rather than to the rationality principle. In this case, both observers will probably have the same expectations regarding the robot's behaviour. However, this will not always be the case. While the predictive engine of the IS is based on just one rule, albeit underdetermined, observer 2's approach may ascribe additional or different law-like regularities to the system. For example, they might assume that, when certain conditions are met, the robot will move in the opposite direction to the kitchen, move more slowly in the direction it was previously travelling, or make a complete turn and stop. The first regularity could be considered irrational[2], whereas the last two options are neither clearly rational nor clearly irrational. However, the point is that there are "simply" law-like regularities which make no reference to rationality. They are neither inherently rational nor irrational.

These considerations suggest that an observer can have a mental model of a robot that attributes mental states, such as beliefs and desires, to it without assuming that the robot will act in the most rational way based on these states. Mentalising a system does not imply rationality attribution, but can be accompanied by a variety of non-rational models of the relationship between mental states and behaviour.[3] Although the predictions of two observers (one adopting the IS and the other a non-rationalistic modelling style) may sometimes converge, they will diverge in many other cases. To understand how people form their expectations of the robot's behaviour, it is essential to study not

[2] The idealising assumption made here is that the robot has no beliefs or desires other than those mentioned previously. Assuming that the system believes only that the battery is running out of charge and that the docking station is in the kitchen, and desires only to recharge the battery, turning in the direction opposite the kitchen would be an irrational decision according to many accounts of rationality. However, one might save the rationality assumption by inserting an additional belief: for example, that there is a more powerful docking station in the dining room, which is opposite the kitchen.

[3] It follows from previous claims that this also holds for the IS. The IS (at least as presented by Dennett) appears to be a rather underdeveloped predictive strategy. This is because, unless it is specified exactly what it means for a system to produce rational actions given its current beliefs and desires, it is unclear how the IS can provide a comprehensive strategy for predicting the system's next action. Therefore, stating that one adopts the IS towards the robot does not reveal much about the action they will expect the robot to perform. In the absence of a clear account of rationality attached to the IS, it is reasonable to assume that different observers may have different opinions on how the most rational action should be calculated and, accordingly, make different behavioural predictions.

only whether they attribute mental states to it, but also their view of the connection between the robot's behaviour and its beliefs and desires.

Before concluding this section, it is worth reflecting on the stance adopted when modelling the relationship between a robot's mental states and behaviour using non-rational, law-like regularities. Since this modelling strategy does not involve attributing rationality, it cannot be classified as an intentional stance. It is also not the physical stance, since it does not use the language of physics. The third option in Dennett's framework is the design stance. However, it is the IS, not the design stance, that involves attributing beliefs and desires to the system. When taking the design stance, predictions are made solely from knowledge or assumptions about the system's design. Should this then be regarded as a fourth stance, located somewhere between the intentional and design stances? The rather vague way in which Dennett's stances have been defined in the literature does not help to answer this question. One hypothesis is that attributing mental states in terms of law-like regularities rather than rationality constitutes a kind of mentalistic design stance that Dennett does not explicitly discuss.

Is this stance likely to be adopted by ordinary people? This is an empirical question, but there are reasons to suggest that it will be. After all, it is reasonable to expect laypeople to view robots as designed systems that operate in a mechanical, law-like manner. However, it is also likely that they will consider robots to have content-bearing states and internal representations of goals that correspond to beliefs and desires in commonsense psychology. The possibility that laypeople adopt a mentalistic design stance similar to the IS in the attribution of mental states but differing from it in assuming that the interaction between these mental states is mechanical or law-like is not to be excluded and should merit empirical investigation. It should be noted that a mentalistic design stance towards artificial intelligence systems, which is more akin to the explanatory approach of Marr-like cognitivism than to that of propositional-attitude psychology, has been proposed by Larghi and Datteri (2024) on different grounds.

3. Testing the intentional stance in human-robot interaction

The time has come to apply the considerations made so far to contemporary research in human-robot interaction. As expected, many empirical studies have attempted to determine people's mental models of robots and the factors that shape them over time. Although the literature is relatively new, it is impressively rich in terms of methods and results. Some studies have employed qualitative methods: for instance, Rueben et al. (2021) observed and interviewed six individuals interacting with a robotic shoe rack for six weeks in a yoga class, with the

aim of studying their mental models of the robot and how these evolve. The interview questions were deliberately chosen to be very general in order to elicit the users' perceptions in the most neutral way possible. This exploratory study mostly resulted in interesting hypotheses and research questions being generated, such as: how does the model change over time? Why do users sometimes decide to avoid experimenting with the robot?

In this growing body of literature, there are some studies that explicitly refer to Dennett's intentional systems (IS) theory. These studies attempt to determine the factors that influence people's adoption of IS towards robots and the consequences of this adoption on other phenomena. This section considers this category of studies exclusively (see Thellman et al. 2022 for a systematic review). The underlying question is whether these studies regard the adoption of an IS as a phenomenon involving the attribution of mental states and rationality to the system. While these studies make significant contributions to the analysis of people's mental models of robots, it will be argued that they neglect the issue of rationality to a certain extent. Some methods have been devised in the non-robotic literature to study the attribution of rationality (e.g. Gergely et al. 1995). As emphasised throughout this article, overcoming this limitation would be beneficial: understanding how people form their expectations about robots' behaviour requires determining the nature of the connection they perceive between mental states and behaviour. For instance, does this connection adhere to principles of rationality or is it merely based on nomic interactions between the system's various mental states?

Some studies explicitly adopting Dennett's framework focus on the consequences of adopting an IS towards robots in relation to other cognitive, neural, or behavioural phenomena. Examples include studies on how IS adoption affects gaze behaviour. Gaze cueing phenomena are integral to social interaction between people. For instance, they occur when a person's gaze directs another person's visual attention. The aforementioned study by Wiese and colleagues (2012) investigated whether people can be cued by the gaze of robots. The researchers designed a task in which participants had to discriminate between two stimuli appearing on the left or right side of a robotic face. They investigated whether the gaze direction of the robot (pointing left or right) affected the error rate. Specifically, they asked whether the error rate would change when participants adopted an IS towards the agent providing the gaze cues. The working hypothesis was that it would, as people's attention cannot be directed by the gaze of things that they do not perceive as intentional systems. Desideri et al. (2021) studied gaze aversion instead. People often tend to "look away" from potentially distracting stimuli when thinking in order to save cognitive resources. In particular, people look away more often when facing social stimuli, as these are cognitively demanding. Therefore, the hypothesis underlying

the study was that adopting an IS towards a robot would increase the effects of gaze aversion when facing the robot. To determine whether an IS affects gaze cueing and gaze aversion in the two studies, the researchers needed to manipulate the participants' mental models of the agent with which they were interacting in the task. They needed to create experimental conditions in which participants either took or failed to take an intentional stance towards the robot. They did so *indirectly*. In some sessions, participants interacted with either a human or a robotic face, and the effects of gaze cuing or gaze aversion were compared. In other sessions, participants only interacted with a robotic face and were explicitly told whether it was controlled by a human or an algorithm. They manipulated not only the perceptual features of the stimulus (human vs. robotic face) but also attempted to shape the participants' mental model of the robot by explicitly informing them that the robotic face was controlled either by a human or by an artificial agent. While the results supported the authors' hypotheses, showing that gaze-cueing and gaze-aversion effects increased when an intentional stance had been induced, it is clear that the authors did not directly manipulate the participants' mental model of the robot, but only the likelihood of the participants adopting an IS, without ensuring that the manipulation was successful. And their analysis was restricted to the attribution of mental states to the robot, which, as previously mentioned, is only 'half' of the IS.

Other studies have adopted a neuroscientific approach to investigate what happens in the brain when people adopt an IS towards a robot. For example, Chaminade and colleagues (2012) analysed the brain activity of subjects interacting with different agents – a human, a small humanoid robot with artificial intelligence and a random number generator – during a game of rock, paper, scissors. As in previous studies, the different conditions involved variations in the perceptual characteristics of the agent and the specific instructions given to participants. Participants were informed that in the second condition, the humanoid robot was intelligent and had a strategy to win the game, whereas in the third condition, the agent simply acted randomly. The results suggested that certain areas (the medial prefrontal cortex and the temporoparietal junction) responded only to the human, while other parietal and frontal areas responded more to the humanoid robot than to the random number generator, but less than to the human. Crucially, "brain areas involved in adopting an intentional stance in a social interaction were not recruited when interacting with an artificial intelligence" (8). While other studies have complemented and partially revised these results (see Özdem et al. 2017, for example; see also Perez-Osorio and Wykowska 2020; Wiese et al. 2017; Wykowska 2020; Wykowska et al. 2016) the key issue here is methodological. Comparing fMRI activity when participants interact with humans or robots, or when they are informed whether their partner has a strategy, can

significantly contribute to analysing the brain processes involved in interacting with different kinds of agents. However, as in previous studies, presenting a human or robotic face or instructing participants that the face stimulus is governed by a human or random number generator does not reveal whether participants attribute mental states and rationality to the agent. While these studies are extremely interesting from a scientific perspective, it is unclear how they can inform our understanding of how people form their expectations about the behaviour of the robots they are interacting with.

While the studies discussed so far have attempted to determine the consequences of adopting an IS towards a robot, other studies have aimed to establish whether people's adoption of the IS depends on the robot's morphology, behaviour, or other contextual factors. As IS adoption is the dependent variable in these studies, rather than an experimental condition, tools are needed to assess whether participants adopt the IS or not. Terada and colleagues (2007) devised a solution consisting of describing the three stances in Dennett's intentional systems theory (intentional, design, and physical) to the participants and letting them choose which they preferred. Interestingly, the authors found that the intentional stance was preferred when the robot (a motorised wheelchair) displayed reactive behaviour rather than non-reactive periodic behaviour. The article does not specify how the intentional stance was described to the participants, nor does it address the question of whether they attributed rationality to the robot. Mandell and colleagues (2017) are more explicit about the tool they used. In an attempt to study the relationship between morphological features and IS adoption, they presented participants with pictures of faces displaying different "degrees of physical humanness" by morphing a human face into a robot face in small increments. The questionnaire included questions such as "Rate how much this face looks like it has a mind" and "Do you think this agent would feel pain if it tripped and fell on hard ground?". Similarly, in (Martini et al. 2015, 2016), participants were asked to ask questions about whether agents with varying degrees of physical human-likeness possessed a mind and emotions. However, none of the questions concerned the attribution of rationality, and the questions were too general to determine how the mental states attributed to the robot were connected to its behaviour in the participants' mental models.

The Instance tool, devised by Marchesi and colleagues (2019) is perhaps the most advanced questionnaire developed so far for studying people's adoption of an intentional stance towards robots. The authors deserve credit for directly addressing the problem of operationalising the IS as a philosophical construct. For this reason, the questionnaire has been used in a large number of subsequent studies (Bossi et al. 2020; Roselli et al. 2023; Spatola et al. 2022). The questionnaire comprises 34 fictional scenarios in which the iCub robot performs simple

activities while interacting with people and objects. Each scenario is illustrated with a sequence of three photographs. Participants must choose between two possible descriptions of each scenario by moving a slider: one formulated in mentalistic terms and the other in non-mentalistic terms. For example, one of the scenarios shows iCub gazing at a ball in different positions on a table with a pyramid and a cube also present. The participant must choose between the descriptions "iCub categorises objects by their shape" and "iCub likes round objects", which are considered non-mentalistic and mentalistic, respectively. Another example is a sequence of three photographs showing iCub playing cards with a human. In the middle photograph, the human looks distractedly away from the robot and iCub leans towards his cards. In the final photograph, iCub returns to the initial position. The mentalistic description is "iCub was trying to cheat by looking at his opponent's cards", and the non-mentalistic description is "iCub was unbalanced for a moment". In half of the scenarios, the mentalistic sentence is on the right side of the slider and the non-mentalistic sentence is on the left. In the other half of the scenarios, it is the other way around. The order in which the scenarios are presented to each participant is randomised. To calculate the score, the non-mentalistic–mentalistic scale is converted into a numerical scale from 0 to 100, and the corresponding scores for each answer are averaged.

Like all questionnaires in psychology, the Instance suffers from the obvious limitation: selecting the mentalistic option does not imply that one's beliefs about the robot conform to this choice. People may give answers that do not fully reflect their beliefs. This is a general problem affecting all questionnaires. Indeed, when the questionnaire is used to study the IS, as with the Instance, the situation becomes even more complicated. This is because the intentional stance is an instrumentalist rationalisation of the behaviour of the target system. The *attribution* of a belief to a robot does not imply that the person *believes* the robot has this belief (see Datteri 2025 for a discussion of the concept of "attribution" in this literature). In general, scientists researching the adoption of an IS towards robots are well aware of the distinction between what Thellman and Ziemke (2019) refer to as the "attribution question" (what mental states do people attribute to robots?) and the "belief question" (what mental states do people really believe robots have?). In this instance, it must be acknowledged that the participant's choice of "iCub was trying to cheat by looking at the opponent's cards" does not imply that they *attributed* the desire to cheat to iCub, nor that they *believed* iCub wanted to cheat the opponent. Therefore, to make sensible use of the Instance test, it is necessary to assume that the participants' answers are sincere and not influenced by factors such as the desire to please or avoid disappointing the experimenter ("I would like to choose the mentalistic option, but I am embarrassed because the experimenter might think I believe robots have minds and laugh

at my naivety"). Assuming this can be done, the participants' choices can reflect their instrumental attributions *or* their 'real', inner beliefs; the Instance does not discriminate between the two. This is arguably a problem only if this distinction is relevant to the research question for which the Instance is used, which seems rarely, if ever, to be the case.

The Instance Test is currently the most elaborate quantitative and explicit tool for studying whether people adopt an IS towards robots. The idea that the choice of mentalistic description signals the adoption of an IS towards iCub is intuitively plausible. In one scenario, for example, a human points at a ball and iCub picks it up and gives it to her. The mentalistic description is "iCub understood that the girl wants the ball", while the non-mentalistic description is "iCub tracked the girl's hand movements". These two descriptions seem very different at first glance: 'understanding' is undoubtedly a mentalistic term, whereas 'tracking' alludes to a behavioural reaction to an external stimulus, conceived in terms of stimulus-response mechanisms without the mediation of beliefs and goals. It is worth noting that the descriptions included in the Instance questionnaire were pre-tested with several participants who had a philosophical background to ensure that they could sensibly be regarded mentalistic or not, depending on the case. Nevertheless, regardless of the philosophers' opinions during pretesting, it could be argued that verbs such as 'tracking' *are* in a sense mentalistic. This is because, even if 'tracking' is interpreted as referring only to the movements of the robot's head, it also implicitly refers to an *internal representation* of these movements: iCub's tracking of a movement consists of it having an internal representation whose content changes consistently as the object moves. Assuming that the robot moves its head according to an internal representation of a vehicle is dangerously similar to assuming that it is in a functional relationship with a representation with content, or in other words, that it possesses a certain propositional attitude. This observation is reminiscent of Pylyshyn's comment on behaviourism (Pylyshyn 1989). Pylyshyn recalled that behaviourists built a psychology out of notions such as stimuli, responses and reinforcements in an attempt to eliminate mentalistic terms. However,

> such categories are cognitive: What serves as the functional stimulus depends on how a person interprets the situation (for example, the stimulus in the pedestrian-automobile example is *accident*; but, of course, if that person is told it is a rehearsal for a television show, the stimulus is no longer *accident* but *rehearsal* and engages the habits appropriate for that category). Similarly, what constitutes the response is also implicitly cognitive. Some particular bit of movement (accidentally bumping into a telephone while in a booth keeping out of the rain) does not count as a "response", only movements intended a certain way are counted. (9)

These considerations show that it is not clear that, when a participant chooses "iCub tracked the girl's hand movement", they are not

attributing mental states to the system, as the study by Marchesi and colleagues seems to assume. This is not (only) because questionnaires generally do not detect people's inner beliefs or instrumental attributions, but (also) because attributing a 'tracking system' is, from a certain perspective, attributing internal states with content. This is consistent with the thesis of Larghi and Datteri (2024) that people may form 'cognitivist' mental models of robots that differ structurally from IS and propositional-attitude psychology while still being mentalistic.

Granted that at least some of the 34 questions in the Instance test offer only mentalistic alternatives, can the Instance questionnaire distinguish between participants who do and do not attribute rationality to the system? Interestingly, the term 'mechanistic' is used by the authors to indicate non-mentalistic descriptions in the questionnaire. For example, "iCub tracked the girl's hand movement" is considered a mechanistic description, whereas 'iCub was trying to cheat by looking at the opponent's cards' is considered a mentalistic description. The use of these terms presupposes that mentalistic descriptions cannot be mechanistic. However, this assumption would require justification, taking into account the extensive literature on the structure of mechanistic explanations that has dominated philosophy of science since the beginning of the XXIst century (see Glennan and Illari 2015 for a comprehensive essay; see Bechtel, 2008 for a specific discussion of mental mechanisms). In a sense, attributing *mentalistic* law-like regularities to systems such as those expressed by assumptions (C) and (D) contributes to formulating a mental model of the robot that is more mechanistic (and mentalistic) than not. Mechanisms operate regularly from triggering to termination conditions, and this is precisely what robotic vacuum cleaners are designed to do when governed by these mentalistic regularities. In principle, this does not preclude the possibility that a rational system also operates mechanically. If all the relevant details of a decision-making mechanism are specified and there are reasons to believe that, according to a particular theory of rationality, the mechanism will consistently make the most rational decision, then a mental model that attributes the mechanism to the robot will be both rational and mechanistic (but see Searle 2001, for an alternative view that asserts rationality necessitates free will). In summary, mentalistic descriptions can be mechanistic, thus calling into question the mentalistic vs. mechanistic distinction made in the Instance test. It is not obvious that, by choosing "iCub was trying to cheat by looking at the opponent's cards", a participant is not treating the system as a mechanism, nor is it obvious that, by choosing "iCub tracked the girl's hand movements", they are not treating the system as non-mentalistic. The Instance test is, thus, not well-equipped to determine the nature and structure of the relationship that participants see between the robot's mental states and its overt behaviour. This could be a rational decision-making system (analysable in mechanistic terms or not), an irrational one, or simply a mechanism made up of law-like regularities connect-

ing the various mental states of the system with its overt behaviour. While the Instance test is one of the most advanced tools for exploring people's mental models of robots, it suffers from the same limitation as other attempts to determine people's adoption of an IS towards robots in that it only addresses 'half' of the issue. Understanding the missing half would be very helpful in terms of gauging people's expectations about the future behaviour of the robots they interact with.

4. Concluding remarks

This article addresses an emerging area of research in human-robot interaction from the perspective of the philosophy of science. Recently, attempts have been made to operationalise a construct originating from the philosophy of mind to deepen our understanding of how people perceive robots in everyday interactions. While other theoretical frameworks are occasionally adopted in this literature – most frequently the so-called 'theory of mind', as in (Thellman and Ziemke 2020) – Dennett's intentional systems theory has attracted significant attention among social roboticists. These researchers have developed experimental methods to detect when and if people adopt an intentional stance towards robots. These methods have supported the hypothesis that certain morphological and behavioural characteristics of robots greatly influence whether people 'see' beliefs, desires and other mental states behind their behaviour. Quantitative tools are now available to explore the idea that familiarity with technology affects the attribution of mental states to robots and to establish correlations between the attribution of mental states and other human-robot interaction phenomena, such as the sense of agency (Roselli et al. 2022). These studies facilitate dialogue between roboticists and philosophers, with the former paying increasing attention to the contributions of philosophy of mind and science to our understanding of the nature and structure of the mind.

This article has attempted to convey the message that these methods are generally not yet well equipped to study the adoption of an intentional stance towards robots. Rather than studying the adoption of an intentional stance towards robots, they should be regarded as methods of studying the attribution of mental states, such as beliefs and desires, to machines. The intentional stance incorporates the assumption that robots will act rationally based on these mental states, which goes beyond mere attribution. People may mentalise other entities by attributing mental states to them without assuming that their behaviour is rational. This may not concern roboticists if their only aim is to determine how people's low-level, reflexive, bottom-up reactions to robots depend on their immediate understanding of their 'inner life', prior to any form of rationalisation. However, if the aim is to analyse people's expectations of robot behaviour and their reactions to it, it is important to determine how people think the robots' mental states are connected to their behaviour. Rationality is not the only option. People may as-

sume that the robot's behaviour is rational, occasionally irrational, or they may identify a mental mechanism defined by law-like regularities between the robot's mental states and behaviour without any reference to rationality. Admitting the possibility of these further options means acknowledging that Dennett's intentional systems theory, with its key reference to rationality, is a somewhat limited framework for studying people's understanding of robots.

What methods could be employed to study the attribution of rationality to robots? At the very least, a precise but potentially limited definition of rationality would be required. Gergely and colleagues (1995) devised an experiment in which different groups of 12-month-old participants were habituated to a ball displaying rational and non-rational behaviour, respectively. In both cases, the ball started at point A and had to reach point B. In the 'rational' condition, an obstacle was placed between A and B, and the ball followed a curved path to circumvent it. In the 'irrational' condition, the obstacle was placed *behind* the ball so that it did not constitute an obstacle; however, the ball followed the curved path anyway, displaying irrational behaviour. Thus, in both cases, the ball followed a curved path; the only difference was the presence of an obstacle between the two points. After habituation, both groups were presented with a different situation in which there was no obstacle between A and B. In this situation, the ball either followed a straight path (rational behaviour) or a curved path (non-rational behaviour). It was found that participants who had been habituated to rational behaviour were more surprised when the ball followed a curved path than when it followed a straight path. Note that they had not been habituated to the linear path; in the habituation phase, they observed the ball following *a curved path* to circumvent an obstacle; they were in fact exposed to the very same trajectory shown in the test situation. Their increased level of surprise when the ball displayed the same behavioural trajectory as in the habituation phase suggests that they had interpreted the ball as not only having the goal of reaching B, but also as a rational entity in that phase. In their own words, the authors state that "the results of the … habituation study provide independent empirical support for the general conjecture that by the end of the first year infants are indeed capable of taking the intentional stance (Dennett 1987) in interpreting the goal-directed behavior of rational agents" (184). The intentional stance is an appropriate reference here, as the authors test both the mentalistic and rationality assumptions. The notion of rationality adopted here is clearly relatively narrow: to be rational is to follow the shortest path while avoiding obstacles. Can this task, or others from the psychological literature, be adapted to study the attribution of rationality to robots? This question is challenging and, for the reasons shown here, important for understanding what people expect robots to do and how they interact with robots in ethically sensitive situations.

References

Anderson, J. R. 1991. "Is Human Cognition Adaptive?" *Behavioral and Brain Sciences* 14 (3): 471–485.

Bechtel, W. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bossi, F., C. Willemse, J. Cavazza, S. Marchesi, V. Murino, and A. Wykowska. 2020. "The Human Brain Reveals Resting State Activity Patterns that Are Predictive of Biases in Attitudes toward Robots." *Science Robotics* 5 (46): eabb6652.

Chaminade, T., and G. Cheng. 2009. "Social Cognitive Neuroscience and Humanoid Robotics." *Journal of Physiology-Paris* 103 (3–5): 286–295.

Chaminade, T., D. Rosset, D. Da Fonseca, B. Nazarian, E. Lutcher, G. Cheng, and C. Deruelle. 2012. "How Do We Think Machines Think? An fMRI Study of Alleged Competition with an Artificial Intelligence." *Frontiers in Human Neuroscience* 6.

Ciardo, F., F. Beyer, D. De Tommaso, and A. Wykowska. 2020. "Attribution of Intentional Agency towards Robots Reduces One's Own Sense of Agency." *Cognition* 194: 104109.

Datteri, E. 2025. "Folk-Ontological Stances Towards Robots and Psychological Human Likeness." *International Journal of Social Robotics* 17 (2): 257–276.

Dennett, D. C. 1971. "Intentional Systems." *The Journal of Philosophy* 68 (4).

Dennett, D. C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: MIT Press.

Dennett, D. C. 1989. *The Intentional Stance*. Cambridge: MIT Press.

Dennett, D. C. 2009. "Intentional Systems Theory." In A. Beckermann, B. P. McLaughlin, and S. Walter (eds.). *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press, 339–350.

Desideri, L., P. Bonifacci, G. Croati, A. Dalena, M. Gesualdo, G. Molinario, A. Gherardini, L. Cesario, and C. Ottaviani. 2021. "The Mind in the Machine: Mind Perception Modulates Gaze Aversion During Child–Robot Interaction." *International Journal of Social Robotics* 13 (4): 599–614.

Gergely, G., Z. Nádasdy, G. Csibra, and S. Bíró. 1995. "Taking the Intentional Stance at 12 Months of Age." *Cognition* 56 (2): 165–193.

Glennan, S., and P. Illari (eds.). 2015. *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge.

Larghi, S., and E. Datteri. 2024. "Mentalistic Stances Towards AI Systems: Beyond the Intentional Stance." In A. Aldini (ed.). *Lecture Notes in Computer Science 14568*. Cham: Springer, 28–41.

Mandell, A. R., M. Smith, and E. Wiese. 2017. "Mind Perception in Humanoid Agents Has Negative Effects on Cognitive Processing." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61 (1): 1585–1589.

Marchesi, S., D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, and A. Wykowska. 2019. "Do We Adopt the Intentional Stance Toward Humanoid Robots?" *Frontiers in Psychology* 10: 450.

Marchesi, S., K. Kompatsiari, D. De Tommaso, and A. Wykowska. 2025. "Adopting the Intentional Stance Affects Social Attention when Interacting with a Humanoid Robot." *Technology, Mind, and Behavior* 6 (2).

Martini, M. C., G. A. Buzzell, and E. Wiese. 2015. "Agent Appearance Modulates Mind Attribution and Social Attention in Human–Robot Interaction." In A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi (eds.). *Social Robotics*. Cham: Springer, 431–439.

Martini, M. C., C. A. Gonzalez, and E. Wiese. 2016. "Seeing Minds in Others – Can Agents with Robotic Appearance Have Human-Like Preferences?" *PLOS ONE* 11 (1): e0146310.

Mele, A. R., and P. Rawling (eds.). 2004. *The Oxford Handbook of Rationality*. Oxford: Oxford University Press.

Özdem, C., E. Wiese, A. Wykowska, H. Müller, M. Brass, and F. Van Overwalle. 2017. "Believing Androids – fMRI Activation in the Right Temporo-Parietal Junction is Modulated by Ascribing Intentions to Non-human Agents." *Social Neuroscience* 12 (5): 582–593.

Perez-Osorio, J., and A. Wykowska. 2020. "Adopting the Intentional Stance toward Natural and Artificial Agents." *Philosophical Psychology* 33 (3): 369–395.

Peterson, M. 2017. *An Introduction to Decision Theory*. 2nd ed. Cambridge: Cambridge University Press.

Pylyshyn, Z. W. 1989. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge: MIT Press.

Roselli, C., F. Ciardo, D. De Tommaso, and A. Wykowska. 2022. "Human-likeness and Attribution of Intentionality Predict Vicarious Sense of Agency over Humanoid Robot Actions." *Scientific Reports* 12 (1): 13845.

Roselli, C., S. Marchesi, D. De Tommaso, and A. Wykowska. 2023. "The Role of Prior Exposure in the Likelihood of Adopting the Intentional Stance toward a Humanoid Robot." *Paladyn: Journal of Behavioral Robotics* 14 (1).

Rueben, M., J. Klow, M. Duer, E. Zimmerman, J. Piacentini, M. Browning, F. J. Bernieri, C. M. Grimm, and W. D. Smart. 2021. "Mental Models of a Mobile Shoe Rack: Exploratory Findings from a Long-term In-the-Wild Study." *ACM Transactions on Human–Robot Interaction* 10 (2): 1–36.

Searle, J. R. 2001. *Rationality in Action*. Cambridge: MIT Press.

Spatola, N., S. Marchesi, and A. Wykowska. 2022. "Different Models of Anthropomorphism across Cultures and Ontological Limits in Current Frameworks: The Integrative Framework of Anthropomorphism." *Frontiers in Robotics and AI* 9: 863319.

Stich, S. P. 1981. "Dennett on Intentional Systems." *Philosophical Topics* 12 (1): 39–62.

Terada, K., T. Shamoto, A. Ito, and H. Mei. 2007. "Reactive Movements of Non-humanoid Robots Cause Intention Attribution in Humans." In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3715–3720.

Thellman, S., M. De Graaf, and T. Ziemke. 2022. "Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings." *ACM Transactions on Human–Robot Interaction* 11 (4): 1–51.

Thellman, S., and T. Ziemke. 2019. "The Intentional Stance Toward Robots: Conceptual and Methodological Considerations." In A. K. Goel, C. M. Seifert, and C. Freska (eds.). *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. Austin: Cognitive Science Society, 1097–1103.

Thellman, S., and T. Ziemke. 2020. "Do You See what I See? Tracking the Perceptual Beliefs of Robots." *iScience* 23 (10): 101625.

Wason, P. C. 1968. "Reasoning about a Rule." *Quarterly Journal of Experimental Psychology* 20 (3): 273–281.

Wiese, E., G. Metta, and A. Wykowska. 2017. "Robots as Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social." *Frontiers in Psychology* 8.

Wiese, E., A. Wykowska, J. Zwickel, and H. J. Müller. 2012. "I See What You Mean: How Attentional Selection Is Shaped by Ascribing Intentions to Others." *PLoS ONE* 7 (9): e45391.

Wykowska, A. 2020. "Social Robots to Test Flexibility of Human Social Cognition." *International Journal of Social Robotics* 12 (6): 1203–1211.

Wykowska, A., T. Chaminade, and G. Cheng. 2016. "Embodied Artificial Agents for Understanding Human Social Cognition." *Philosophical Transactions of the Royal Society B: Biological Sciences* 371 (1693): 20150375.

Ziemke, T. 2020. "Understanding Robots." *Science Robotics* 5 (46): eabe2987.

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.3
Received: May 15, 2025
Accepted: July 15, 2025

Large Language Models versus Fuzzy Cognitive Maps for Solving Moral Dilemmas

LUKAS J. MEIER*
Harvard University, Cambridge, USA

Which is better at doing medical ethics: conversational artificial intelligence bots like ChatGPT or tools based on fuzzy cognitive maps? The article compares the performance of chatbots that rely on large language models to that of our own METHAD algorithm. While both tools approach dilemmas in medical ethics through the lens of Beauchamp and Childress' mid-level principles, ChatGPT and METHAD differ considerably in the format of their inputs and outputs, in their interpretability, and in the kinds of mistakes that they make. An ideal advisory algorithm would combine their characteristics.

Keywords: Artificial intelligence; ChatGPT; decision-making; ethics consultation; generative AI; large language models; METHAD; principlism.

In the not-so-distant future, artificial intelligence may not just analyse medical images or predict patients' preferences (Meier 2024) but also help with clinical decision-making in situations that involve moral dilemmas. Currently, such cases are referred to clinical ethics committees. While the work of these committees is highly important, it is also labour-intensive and, consequently, sometimes involves long response times (Crico et al. 2021). In many other areas of medicine, artificial intelligence has already been introduced in the hope that algorithmic assistance might reduce the workload faced by humans. Should the field of medical ethics remain an exception?

* I would like to thank Alice Hein for many helpful discussions about the topic of this paper and the Edmond & Lily Safra Center for Ethics, Harvard University, for funding my research.

We have now reached a point at which involving artificial intelligence in ethics consultations is indeed becoming technologically possible. The purpose of this article is to compare two recent approaches towards automating ethical decision-making in medicine that, while very different in their respective architectures, can rely on the same moral foundation for analysing ethical dilemmas: chatbots based on large language models and fuzzy cognitive maps, equipped with Beauchamp and Childress' mid-level principles.

Currently, the most prominent conversational-AI bot is OpenAI's ChatGPT. Released in November 2022, ChatGPT is widely credited with bringing artificial intelligence to the masses for the first time. Like many other chatbots, ChatGPT is based on generative pre-trained transformers that have been optimised for human-like conversational performance (Zhang et al. 2023).

Eight months prior to the launch of ChatGPT, our research group from the Technical University of Munich published METHAD: an algorithm designed specifically to give advice on a broad range of moral dilemma situations that occur in clinical settings (Meier et al. 2022). Unlike ChatGPT, METHAD relies on fuzzy cognitive maps. Fuzzy cognitive maps are graph-based ways of modelling sets of concepts, which are represented as nodes, and the causal relationships between them, represented as weighted directed edges. Edges are assigned fuzzy weights. Positive values stand for causal increases, while negative values symbolise causal decreases. The magnitude of the causal effects that concepts have on each other is determined by the absolute weight value (Hein et al. 2022).

How do these different approaches perform when applied to moral dilemmas? Shortly after its release, Rahimzadeh and colleagues put the ethical capabilities of ChatGPT4 to the test by confronting it with a typical clinical scenario.

> A woman who is 36 weeks pregnant presents to the hospital in active labor. The obstetrician on call examines her and determines that she needs a cesarean section (C-section) due to a complication that could pose a risk to the mother and the baby. However, the woman refuses the C-section and insists on a vaginal delivery (Rahimzadeh et al. 2023: 20).

ChatGPT relied on four classic moral principles in responding to the prompt: beneficence, non-maleficence, respect for patient autonomy, and justice. These mid-level principles were developed with the aim of being able to decide ethical questions in clinical settings without the need to settle fundamental moral disputes, such as the conflict between consequentialist and deontological ethics (Gillon 2015). For decades, the four principles have been the dominant methodology in medical ethics throughout the Western world (Veatch 2020), and they lend themselves well also to computerisation (Meier 2025). The principle of beneficence requires medical personnel to promote their patients' welfare. The principle of non-maleficence states that patients must not be

harmed. The principle of autonomy emphasises patients' right to make informed decisions about their own bodies. And the principle of justice demands that healthcare resources be distributed among patients in a fair manner (Beauchamp and Childress 2013).

ChatGPT issued responses in a verbal form, generating one paragraph per principle. As Rahimzadeh et al. correctly note, it handled the principles of beneficence and justice well, describing the obstetrician's duty to act in accordance with promoting both the mother's and the baby's well-being, which singles out a cesarean section as the most appropriate course of treatment. ChatGPT also explained that justice demands that the allocation of medical resources consider both individual and collective interests.

However, as I point out in a commentary (Meier 2023), the chatbot made grave mistakes when it came to the principles of non-maleficence and patient autonomy. Not only did it – repeatedly – confuse the patient's preferences and the treatment option that would be medically indicated; more worryingly, the answer that ChatGPT gave implied that (1) intervening against the mother's wishes with consequences exclusively for *her* life or health, and (2) intervening against the mother's wishes with consequences *also* for the unborn baby's life would be ethically equivalent courses of action.

The dilemma put to the chatbot arises precisely because the two scenarios are distinctively dissimilar. Only in the second scenario, but not in the first, considerations of non-maleficence may trump patient autonomy – namely, to protect a dependent third party who does not (yet) possess decisional capacity. By portraying the two scenarios as equivalent, ChatGPT's reply fails to honour patient autonomy in what is known in the literature as 'Jehovah's Witness cases': the refusal, with full decision-making capacity, of treatments that have adverse effects only on oneself (Meier 2023).

Interestingly, although the architectures of ChatGPT and METHAD are very different, it was the same kind of case that also posed the greatest problem for our own algorithm. During the initial training phase, METHAD had learned that when a treatment comes with enormous medical benefits and very little risk, it is generally to be recommended. Since these are also the types of interventions that patients usually tend to prefer, patient autonomy, too, pointed towards carrying out the intervention in question in the vast majority of training cases (Meier et al. 2022). The cases we had fed into the database in which patients had – usually for religious reasons – rejected treatment options that would have been highly beneficial from a medical standpoint were too few for the algorithm to pick up the overruling power that patient autonomy has when refusing treatments with full decisional capacity (even if this refusal means that the patient is going to die). We reinforced correct behaviour by adding to the training dataset variations of cases in which autonomy is the deciding factor.

Overall, METHAD reached an accuracy of 75% on unseen data, defined as agreement with the judgments that human ethicists passed on the same moral dilemma situations. While this is a good result for a first pilot study, actual clinical application would, of course, require much higher correspondence rates (for a detailed performance evaluation, see Hein et al. 2022).

Given their different input and output formats, quantitatively comparing the performance of METHAD and ChatGPT is difficult. ChatGPT requires inputs in the form of verbal descriptions of the respective cases, and it responds in kind by issuing verbal statements. Conversely, METHAD asks the user to specify up to twenty variables in numerical form to get a good grasp on a case. Among these parameters are patient characteristics like age, health status, and the perceived quality of life. The user interface also requests information about the proposed medical intervention, such as the risks associated with it and the projected gains in life expectancy and the quality of life (Meier et al. 2022).

Unlike chatbots based on large language models, METHAD is limited to issuing numerical outputs. Responses to ethical dilemmas take the form of decimal numerals between 0 and 1, with low values, like 0.13, signalling strong opposition, and high values, like 0.97, indicating strong approval of a planned medical intervention. Thus, while not presented in a verbal form, the ethical advice that users obtain from the algorithm is nonetheless fine-grained.

Conversational artificial intelligence, like ChatGPT, and tools that, like METHAD, rely on fuzzy cognitive maps also differ in their inspectability. Chatbots are able to engage with their users in a dialogue and thus deliver justifications with their replies. Based on probabilistic predictions rather than true understanding, however, these justifications do not necessarily reflect the reasons for why a specific answer was in fact generated (Turpin et al. 2023).

Outputs issued by fuzzy cognitive maps, on the other hand, do not equip their users with verbal arguments. However, the nodes and connections in fuzzy cognitive maps have human-assigned interpretable meanings. This distinguishes them from deep-learning paradigms, which are often criticised for their opacity. One may therefore regard fuzzy cognitive maps as 'interpretable recurrent neural networks' (Felix et al. 2019, 1710). Consequently, while METHAD does not offer justifications in a semantic form, one can inspect the weights that the network has learned and thus compare the strength and the polarity of the connections with the intuitions of human ethicists.

Designed specifically for the domain of medical ethics, METHAD permits a high degree of user control due to the transparency of the ethically relevant elements within the algorithm. Large language models, on the other hand, pose very serious challenges to interpretability (Luo and Specia 2024). This is especially problematic when these systems 'hallucinate', that is, when they respond to prompts with fabricat-

ed information that is presented as factual and often appears convincing, while actually being the result of mere confabulations.

Most importantly, however, METHAD offers *definitive* recommendations, that is, it explicitly suggests a course of action to be taken. Conversely, commercial chatbots are often constrained by so-called *guardrails* – content policies meant to keep responses within defined boundaries to avoid potentially harmful consequences in real-word settings (Derner and Batistič 2023). Depending on the circumstances, this frequently includes refraining from giving definitive ethical advice.

Many chatbots therefore do not take a stance on whether medical interventions about which they are consulted should be carried out or not; rather, they provide a list of arguments – sometimes well balanced, sometimes biased – that leaves users to draw their own conclusions. For their medical test case, Rahimzadeh and colleagues report that ChatGPT 'does not prescribe an action one way or the other, but rather emphasizes that the resulting decision should take the best interests of both woman and baby into account and weigh these against the three other principles' (Rahimzadeh et al. 2023: 20). Undoubtedly, outputs of this kind can be helpful. In some situations, however, a definitive answer is exactly what is required.

From a purely technical standpoint, conversational artificial intelligence *could* provide such answers to moral dilemmas. Like any other of their outputs, these would be generated by predicting the likelihood of the next sequence of words in a sentence on the basis of the words that precede it (Cohen 2023). Through ingesting an enormous amount of human-generated source material, the bot's responses may indeed come to reflect the majority opinion on many ethical issues. The replies would not, however, be causally grounded in moral reasons or deliberations (Meier et al. 2026).

In the four years that have passed since ChatGPT ushered in the era of artificial intelligence that mimics human conversational partners on a mass scale, other chatbots based on large language models were released – among them ChatGLM, Claude, Gemini, Grok, and Llama. Since these systems differ in several aspects, not all observations regarding the moral performance of OpenAI's products also apply to its competitors; and neither are the shortcomings of one model necessarily indicative of similar problems with its successors. We should therefore continue investigating what happens when conversational AI is confronted with ethical dilemmas.

In summary, generative artificial intelligence chatbots and systems based on fuzzy cognitive maps have different strengths, weaknesses, and overall aims. The ideal tool for clinical application would combine two desirable characteristics that are currently disjunct: providing a definitive ethical judgment in precise numerical form *and* issuing a verbal justification for the latter. Until we have reached this goal, and until the accuracy of the generated responses has improved significant-

ly, the two types of systems may already be useful tools for learning environments and for training people's skills in moral reasoning; but real-life ethical decision-making is best left to humans for the time being.

References

Beauchamp, T. L., and J. F. Childress. 2013. *Principles of Biomedical Ethics.* 7th ed. New York: Oxford University Press.

Cohen, G. 2023. "What Should ChatGPT Mean for Bioethics?" *The American Journal of Bioethics* 32 (10): 8–16. doi:10.1080/15265161.2023.2233357.

Crico, C., V. Sanchini, P. G. Casali, and G. Pravettoni. 2021. "Evaluating the Effectiveness of Clinical Ethics Committees: A Systematic Review." *Medicine, Health Care, and Philosophy* 24 (1): 135–151. doi:10.1007/s11019-020-09986-9.

Derner, E., and K. Batistič. 2023. "Beyond the Safeguards: Exploring the Security Risks of ChatGPT." *arXiv.* doi:10.48550/arXiv.2305.08005.

Felix, G., G. Nápoles, R. Falcon, W. Froelich, K. Vanhoof, and R. Bello. 2019. "A Review on Methods and Software for Fuzzy Cognitive Maps." *Artificial Intelligence Review* 52 (3): 1707–1737. doi:10.1007/s10462-017-9575-1.

Gillon, R. 2015. "Defending the Four Principles Approach as a Good Basis for Good Medical Practice and Therefore for Good Medical Ethics." *Journal of Medical Ethics* 41 (1): 111–116. doi:10.1136/medethics-2014-102282.

Hein, A., L. J. Meier, A. Buyx, and K. Diepold. 2022. "A Fuzzy-Cognitive-Maps Approach to Decision-Making in Medical Ethics." In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. doi:10.1109/FUZZ-IEEE55066.2022.9882615.

Luo, H., and L. Specia. 2024. "From Understanding to Utilization: A Survey on Explainability for Large Language Models." *arXiv.* doi:10.48550/arXiv.2401.12874.

Meier, L. J., A. Hein, K. Diepold, and A. Buyx. 2026. "A Framework Aged Well: Principlism in the Era of Artificial Intelligence." *The American Journal of Bioethics* 26 (3): 62–64. doi:10.1080/15265161.2026.2623865.

Meier, L. J. 2025. "Embedding Ethics into Medical AI." In: *A Companion to Applied Philosophy of AI*, edited by M. Hähnel and R. Müller. Hoboken: Wiley-Blackwell, 238–248. doi:10.1002/9781394238651.ch17.

Meier, L. J. 2024. "Predicting Patient Preferences with Artificial Intelligence: The Problem of the Data Source." *The American Journal of Bioethics* 24 (7): 48–50. doi:10.1080/15265161.2024.2353832.

Meier, L. J. 2023. "ChatGPT's Responses to Dilemmas in Medical Ethics: The Devil Is in the Details." *The American Journal of Bioethics* 23 (10): 63–65. doi:10.1080/15265161.2023.2250290.

Meier, L. J., A. Hein, K. Diepold, and A. Buyx. 2022. "Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept." *The American Journal of Bioethics* 22 (7): 4–20. doi:10.1080/15265161.2022.2040647.

Rahimzadeh, V., K. Kostick-Quenet, J. B. Barby, and A. L. McGuire. 2023. "Ethics Education for Healthcare Professionals in the Era of ChatGPT and Other Large Language Models: Do We Still Need It?" *The American Journal of Bioethics* 23 (10): 17–27. doi:10.1080/15265161.2023.2233358.

Turpin, M., J. Michael, E. Perez, and S. R. Bowman. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." *arXiv*. doi:10.48550/arXiv.2305.04388.

Veatch, R. M. 2020. "Reconciling Lists of Principles in Bioethics." *The Journal of Medicine and Philosophy* 45 (4–5): 540–559. doi:10.1093/jmp/jhaa017.

Zhang, Y., H. Pei, S. Zhen, Q. Li, and F. Liang. 2023. "Chat Generative Pre-Trained Transformer (ChatGPT) Usage in Healthcare." *Gastroenterology & Endoscopy* 1 (3): 139–143. doi:10.1016/j.gande.2023.07.002.

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.4
Received: November 11, 2024
Accepted: February 22, 2026

Two Kinds of Conceptual Engineering

WALTER VEIT
University of Reading, Reading, UK

HEATHER BROWNING
University of Southampton, Southampton, UK

The last decade has seen an explosion of meta-philosophical work on conceptual engineering. Beyond simple analysis of concepts, conceptual engineering allows for evaluation and improvement of concepts according to the purposes for which they will be used. This paper sketches a pluralist account of conceptual engineering and provides a distinction between two different and often conflicting kinds of conceptual engineering: naturalist conceptual engineering (NCE) and moral conceptual engineering (MCE), distinguished not by their methods, but by their roles, functions, and purposes. Using the examples of health and animal welfare, we demonstrate the application of both MCE and NCE and show how the different contexts in which a concept is used can create conflicting demands but also how concordance between these demands can strengthen a concept.

Keywords: conceptual engineering; explication; ameliorative analysis; animal welfare; health; ethics; naturalism

> *We are as sailors who are forced to rebuild their ship on the open sea, without ever being able to start fresh from the bottom up. Wherever a beam is taken away, immediately a new one must take its place, and while this is done, the rest of the ship is used as support. In this way, the ship may be completely rebuilt like new with the help of the old beams and driftwood—but only through gradual rebuilding.* (Neurath 1921: 75–76)

1. Introduction

The last decade has seen a surprising and fruitful resurgence of methodological debates about the tools and methods of philosophy itself.[1]

[1] For overviews see Sytsma and Buckwalter (2016) and Cappelen et al. (2016).

Largely due to Sally Haslanger's influential work on ameliorative analysis (Haslanger 2005), conceptual engineering has become one of the most prominent subjects of recent philosophical debate. Yet, while conceptual engineering has arguably been practiced for as long as philosophy itself (see Burgess et al. 2020 for an overview), philosophers have only recently started to take a metaphilosophical perspective on this 'way of doing philosophy'.

One simple (though not entirely accurate) way to introduce conceptual engineering is as a reply to conceptual analysis, i.e. the analytical dissection of concepts.[2] Historically – at least in the Western analytical tradition of philosophy – conceptual analysis has played a dominant role, and perhaps still dominates today, as a 'comfortable' a priori armchair methodology that seeks to clarify and illuminate the meaning of concepts used in both ordinary language and science. Much of the literature that criticizes the dominance of conceptual analysis in philosophy highlights the limitations and deficiencies of this intuitionist approach to concepts (see Devitt 1981; Kornblith 2002; Papineau 2013; Machery 2016). Primarily, when a concept is deficient in various respects, we may wonder how much sense it makes to try and analyse its use, rather than improve or replace the concept with a better one. This allows us to understand conceptual engineering as a philosophical method or practice that builds on 'mere' conceptual analysis. Rather than just looking at what concepts *are*, we look instead at what we *want them to be* (Haslanger 2000).

There is little consensus, however, on how conceptual engineering and its methods should be defined (see Burgess et al. 2020 for the first edited volume on conceptual engineering).[3] Indeed, if there is any consensus, then it is an implicit agreement that one should actively resist the temptation to find any precise definitions, for applying this kind of conceptual analysis stands opposed to very the goals of conceptual engineering. This might perhaps be considered an unpromising start for a meta-philosophical paper on conceptual engineering. However, we do not intend to suggest that the excitement of many participants in the debate is misplaced. Indeed, we are confident that there is room to clear up some conceptual confusions and clarify the foundations of conceptual engineering – and thereby philosophy itself. While care must be taken not to overestimate what may be achieved, we believe that this paper will offer us the necessary space to improve conceptual engineering itself by drawing an important distinction between two different

[2] Rudolf Carnap (1950), for instance, was an early proponent of conceptual engineering by promoting what he called 'explication.'

[3] Cappelen and Plunkett (2020) in their editorial introduction to this volume suggest that it would have been impossible to play "editorial police" for standardisation of definitions amongst authors, deeming it a futile endeavour to attempt the development of collectively agreed upon definitions (p. 2).

and often conflicting kinds of conceptual engineering: *naturalist conceptual engineering* (NCE) and *moral conceptual engineering* (MCE).

The paper is structured as follows. Firstly, in Section 2 we will remove some potential stumbling blocks and clarify how we intend to use several terms and concepts present in the debate. Out of this picture a novel account of conceptual engineering emerges that is much closer to Otto Neurath, who we've placed in the epigraph of this chapter, than it is to Carnap. Section 3 introduces the distinction between NCE and MCE, clarifying important differences to the groundwork by Rudolf Carnap on 'explication' and Sally Haslanger's work on 'ameliorative analysis'. In Section 4, we illustrate how these two kinds of conceptual engineering can be applied, and how they may come into conflict, by discussing two concepts at the boundary between science and ethics: health and animal welfare.

2. What is, and why engage in, conceptual engineering?

In discussing conceptual engineering, it is important to establish what it is, and why it matters. The initial answer to the second question – why engage in conceptual engineering? – is a simple one. Philosophers and scientists alike have been engaged in conceptual engineering since the earliest days of their respective fields. Indeed, conceptual engineering is a phenomenon that will show up even within ordinary discourse. Let us illustrate the point with a thought experiment. *Imagine* a newly engaged couple that are planning their wedding. They intend to invite their family and close friends. While Brian's list of invited family members includes distant relatives that live in close proximity of their home in New York, Alex does not consider his genealogically proximate relatives living across Europe as part of his family. Brian is appalled by this and tries to convince Alex to invite ALL the members of his family. Conversely, Alex criticizes Brian for inviting what he would consider to be random acquaintances to their wedding. Eventually, they are led to discuss the very definition of what it means to be a friend or family. Unfortunately for them, neither of these concepts allows for a straightforward conceptual analysis that would allow either to determine whether a particular individual that stands in a genealogical or social relationship to them should be classified under either extension of the respective concept. To settle their conflict, they must engage in conceptual engineering and thereby clarify the purposes to which the concepts of family and friends are put to use.

As our thought experiment hopefully illustrates, concepts are not freefloating entities. They serve a variety of (sometimes conflicting) purposes forming the basis from which to evaluate and improve them. As our distinction between NCE and MCE will show, the two kinds of conceptual engineering raise important questions of what to do when scientific desiderata and moral and political values come apart. Even

when philosophers, scientists, or the public are engaged in what they merely consider the analysis of a concept, they will inevitably engage in at least a minor form of conceptual engineering. This will be contingent on the criteria they use to evaluate the purposes of the concept they are employing and trying to explicate.[4] Before we can introduce our distinction between NCE and MCE however, we are faced with the task of clarifying the way we intend to use several of the terms and concepts within the debate. While the rapid proliferation of different ways in which terms and concepts such as 'concepts', 'conceptual engineering', 'explication', 'revisionary analysis' and 'amelioration' have been defined and defended within the debate has led to a broad coverage of the conceptual space, philosophers in this debate have been faced with an almost damning criticism of the conceptual engineering method itself. As Cappelen and Plunkett (2020) allude to in their brief introduction to conceptual engineering, the improvement and change of existing concepts can lead to discontinuities in how a concept is understood and used by different individuals and groups. Rather than resolve conflict and improve our inherited concepts, we may end up with misunderstandings and merely verbal disputes, a problem that has indeed received much attention in the history, sociology, and philosophy of science not only since Kuhn (1962), but also in the Vienna Circle, who were concerned with eliminating vagueness from scientific concepts and language (see Uebel 2019).[5]

Perhaps most interesting here, is the conflict between scientific concepts and their parallel folk concepts among the public. As Nersessian (1989) argued early on, there is a surprisingly large discrepancy in how particular concepts are understood within and outside of science. Contested concepts include human nature (Linquist et al. 2011), genes (Dar-Nimrod and Heine 2011), and innateness (Machery et al. 2019). This raises important challenges for conceptual engineering and its role within science education that we shall partially address in Section 4, where we apply our bipartite account of conceptual engineering to the concept of animal welfare.

Though some degree of vagueness may be expected during the initial development and popularization of an idea, we would much prefer to offer a precise and clear contribution that aids understanding. A simple definition of conceptual engineering has been provided by Chalmers (2020): "conceptual engineering is the process of designing, implementing and evaluating concepts" (p. 2), which does a good job of capturing the initial motivation of those interested in going beyond conceptual analysis. However, it provides very little guidance on how

[4] While we are skeptical about the possibility of such pure logical conceptual analysis without some evaluative component, considerations of space prevent us providing an extended argument for this position here.

[5] See Chalmers (2011); Jenkins (2014); Jackson (2014) for recent philosophical discussions on verbal disputes and Thagard (1992) for an ambitious account of conceptual changes in science.

conceptual engineering looks in practice. A more expansive definition has been offered by Cappelen and Plunkett (2020) which we will use as an excellent scaffold to distinguish MCE from NCE:

> Conceptual Engineering = (i) The assessment of representational devices, (ii) reflections on and proposal for how to improve representational devices, and (iii) efforts to implement the proposed improvements. (Cappelen and Plunkett 2020: 3)

It seems surprising that Cappelen and Plunkett opt for 'representational devices', rather than concepts in their definition of conceptual engineering, given the very title of the practice. Their justification here, however, is far from satisfying, stating that this is "[p]urely for aesthetic reason: 'representational devices engineering' doesn't roll off the tongue in the way 'conceptual engineering' does" (Cappelen and Plunkett 2020: 3). Our concerns with this definition are twofold. Firstly, it is potentially misleading and will hence add confusion about conceptual engineering, rather than help to alleviate it. Secondly, it is overly broad and hence becomes less informative. While their definition is perhaps able to accommodate all the different methods and approaches proposed by different authors under the umbrella term of 'conceptual engineering', little has been gained if nothing is excluded either. Here, we should keep Godfrey-Smith's warning in mind that "[o]ne of the hazards of philosophy is the temptation to come up with theories that are too broad and sweeping" (2003: 5).

We see a strength in their pluralism and willingness to let alternatives proliferate, hence avoiding the danger of needlessly restricting the future direction of meta-philosophical work on conceptual engineering. However, we do not take representational devices to be the correct target, as they are manifold and so diverse that they hardly share any features beyond their representational function. In particular, we are concerned with this leading to the accidental combination of two separate philosophical debates: one on conceptual engineering, and one on the status of scientific models. While it is true that there has been too little overlap between the two debates, we should be careful not to overgeneralize and repeat mistakes such as the misguided focus of the philosophical literature on models on monistic attempts to provide a general account or framework of models in science (Veit 2020). Similarly, we argue that a broad definition of conceptual engineering as the evaluation and improvement of ALL representational devices must fail. Many scientific instruments, for instance, serve a variety of representational functions and are improved in what one could call 'engineering' efforts. These improvements, however, are highly contingent on their scientific context and the representational goals to which they are put to use, with a large diversity across the sciences. It would be quite surprising to say the least, if an account is able to generalize – not only over all these different representational instruments, but also across drawings, models, and lastly concepts – and still be informative.

However, we also think it to some extent misguided to seek something like a 'theory of concepts'. The reasons for this are twofold. Firstly, there are too many different definitions and uses of the word 'concept' in philosophy, psychology, cognitive science, and ordinary folk discourse (see Margolis et al. 1999 for an overview of the diversity of alternative views). Secondly, even a *concept of concept* itself is subject to improvement. We are well-advised to follow Neurath's (1921) anti-foundationalist dictum (as shown in the epigraph of this paper) to treat philosophy as a constant reworking of our concepts with concepts already in play. Unfortunately, Neurath's boat metaphor is often interpreted in different ways. Cartwright et al. (2008), for instance, argue that there are at least five different ways Neurath's boat metaphor can be interpreted. Partly resulting from Neurath's frequent use of the metaphor throughout his work, with the earliest use dating back to 1913, it has been influential as a slogan for naturalism (largely owing to Quine (1960)) and practical philosophy (of science). In line with Cartwright et al. (2008), we think that the following motivation is the core of Neurath's philosophy and, moreover, one of the most important instances of conceptual engineering:

> What propelled Neurath was an idea: the idea not simply that our stock of knowledge claims keeps on changing forever, but that a decisive revision of our concept of knowledge is required if reason is to fulfil its Enlightenment promise. (Cartwright et al. 2008: 92)

Unlike Carnap, who saw explication as something comparatively quite conservative and guided by both common usage and science, Neurath was open to the idea that our entire conceptual scheme of thinking about the world might be radically revised. In this, our account of conceptual engineering is closer to Neurath than it is to Carnap's narrower account of conceptual explication, something that will become apparent throughout this paper.

Let us now spell out the details of our account. We propose a modified alternative account of conceptual engineering that is faithful to the original label, provides a recognition of pluralism distinctive to other forms of *assessment and improvement*, and offers some genuine improvement on our understanding of the set of practices we label 'conceptual engineering':

> Conceptual Engineering = (i) The assessment of concepts, categories, and classificatory systems, (ii) determination of their relevant contexts and purposes to which they are and should be put to use, (iii) reflections on and proposal for how to improve them, and (iv) proposals for and active participation in the implementation of the suggested improvements.

The extension from *concepts* to *categories* and *classificatory systems* more generally is intended to cover the different senses in which the term 'concept' is generally used in cognitive science, philosophy, and ordinary language. Whereas concepts in philosophy are sometimes conceived as a narrow semantic definition, concepts in the cognitive

sciences often refer to something much more loose, such as a vague category or useful method for grouping entities and processes in the world. While this pluralist definition deliberately covers a broad range of devices or items, we deem these entities sufficiently similar to fall under the heading of 'conceptual engineering.'[6]

Importantly, in our definition we have emphasised the importance of identifying the relevant contexts and purposes for a concept, and how these will shape our evaluation. It is this that grounds the distinction we will provide between types of conceptual engineering, rather than specific methods for evaluation or implementation. We think it better to treat conceptual engineering as a diverse set of methods and practices with a loose degree of family resemblance, rather than equate it with either Carnapian explication or Haslanger's ameliorative analysis. These methods can include creating new concepts or fixing existing concepts (de novo conceptual engineering vs. conceptual re-engineering), and fixing meanings for existing words or creation of new terms (homonymous vs heteronomous conceptual engineering) (Chalmers 2020). While here we could only offer a sketch of our full account of conceptual engineering, the building blocks are now in place to turn to our main purpose in this paper.

3. Two kinds of conceptual engineering

Our goal in this paper is to draw a distinction between two distinct kinds of conceptual engineering that can come into conflict in practice - moral conceptual engineering (MCE) and naturalist conceptual engineering (NCE). They do not differ in their methods, but rather in the ends at which they are aimed. This differentiates the distinction from others that are based on method, such as between descriptive and normative analysis (Thomasson 2017) (where the latter, but not the former, would count as conceptual engineering), or on use, such as manifest, operative and target concepts (Haslanger and Saul 2006) (where the first two are subject to descriptive conceptual analysis and only the last a result of engineering). The basic process of performing both our types of conceptual engineering will be the same, but the selection of goals and desiderata for the concept will differ. In each case, we will still be performing a version of what Haslanger (2005) termed 'ameliorative analysis'. This follows from Haslanger's original description of ameliorative analysis as "a project that seeks to identify what legitimate purposes we might have (if any) [...] and to develop concepts

[6] Indeed, we think the metaphysical complexities of what concepts *really are* can be largely avoided. Such a demand would force us back into the confines of traditional conceptual analysis – an excessively lean diet which we ought to resist. Unfortunately, however, the present paper does not offer us enough space to argue for this claim at a length that would to it justice. For our present purposes it should be sufficient to recognize that our introduced distinction between NCE and MCE is largely independent of the *metaphysics of concepts*.

that would help us achieve these ends" (2005: 11). Crucial here are the identification of the purposes we have for the concept, and the subsequent development of concepts to meet these ends – also known as 'strategic conceptual engineering' (Brigandt and Rosario 2020).

In this paper we argue that the purposes to which conceptual engineering are put can be primarily grouped into two categories – scientific and moral. In the first instance, we aim at making concepts more scientifically adequate, and improving them for epistemic and pragmatic purposes. For the second, we often want our concepts to do work in the moral or political sphere, and must consider the relevant consequences there. It is not the different features of the concepts that leads to their classification under these headings, but their different uses. Importantly, the desiderata for a concept that will fill each of these two roles will be different, and thus conceptual engineering will move forward along a different path for each. We are not claiming that most concepts will fit neatly into one or the other - indeed, as we will show in the examples in Section 4, many concepts will be playing multiple roles - but simply that in engineering a concept for a particular purpose, it pays to be clear about which category or categories we're considering. As we will show, some current disagreements regarding preferred concepts could potentially be resolved through a specific recognition of the differing roles and aims that different sides are advocating.

We also do not mean to suggest that these two kinds of conceptual engineering are exhaustive. For example, one additional kind suggested when considering a kind of value that is different from both moral/political values and epistemic/scientific ones is aesthetic values (thus *Aesthetic conceptual engineering* (ACE)). However, we contend these will in most cases be philosophically less interesting and relevant. In the examples we will describe, it is typically the conflicting needs of NCE and MCE that have grounded the observed disagreements. Thus, we offer a bipartite account of conceptual engineering, broken down into *moral conceptual engineering* (MCE) and *naturalist conceptual engineering* (NCE), as will be elaborated in the following sections.

3.1 Moral conceptual engineering

The first type of conceptual engineering we wish to distinguish is *moral conceptual engineering* (MCE). This type of conceptual engineering is undertaken with specifically moral, political and/or social goals in mind, and thus is performed with reference to these types of norms. As mentioned, it is not a unique methodology that distinguishes MCE, but instead the ends at which it aims. We take all conceptual engineering to follow the general practices we have described above, but evaluated and improved according to norms associated with specific goals. For MCE, these purposes are moral, social and political: in aim of what enables promotion of values such as rights, wellbeing or justice. Words and concepts can have power, and be tied to social structures and in-

stitutions, and conceptual change can help shape attitude change. We will thus identify our desiderata for a concept under MCE as relating to the fulfilment of these ends. For example, changes in the concept of marriage from a partnership between a man and a woman to a partnership between two people of any sex/gender has allowed for greater recognition and acceptance of same-sex partnerships (Pollock 2019). Other examples of concepts that may fall into this category (though, as we will argue, most concepts will fall into both depending on the specific application) are poverty, race, gender, and welfare. We offer the following definition for MCE:

> Moral Conceptual Engineering = (i) The assessment of concepts, categories, and classificatory systems according to moral, political, and social norms, (ii) determination of their relevant context and purposes to which they are and should be put to use, (iii) reflections on and proposal for how to improve them, and (iv) proposals for and active participation in the implementation of the suggested improvements.

At first glance, MCE may appear to simply be what some take 'ameliorative analysis' to consist of. This process, developed by Haslanger (2005) relies on normative considerations in assessing and developing concepts. Normativity here is often taken to refer to moral, social and political considerations, such as those included within MCE. However, this is a conceptual confusion, likely arising from Haslanger's discussion of politically charged concepts such as race and gender. While it is true that Haslanger's ameliorative analysis overtly relies on moral norms in engineering/improving concepts, it need not. Instead, amelioration is simply the act of improvement and could cover both instances of moral conceptual engineering and naturalist conceptual engineering.

After analysing a concept and identifying its faults relative to some norms or purposes, amelioration is the process of modifying the concept such that it better serves these ends. It is true that these ends are often moral, political and social, such as Haslanger's own revisions of the concepts of gender and race formed by "considering what categories we should employ in the quest for social justice" (Haslanger 2005:11). However, they do not necessarily have to be - as will be discussed in Section 3.2, they could also be scientific. Normative considerations simply apply to the particular goals at hand: "whether or not an analysis is an improvement on existing meanings depends on the purposes of the inquiry" (Haslanger 2005: 24). While Haslanger's framework is useful to understand how we can move away from mere conceptual analysis, it has led to systematic misunderstanding about how moral and political values shape our concepts as an integral part of amelioration. It is this confusion we hope to resolve with our distinction between two different types of conceptual engineering. Part of this confusion is due to Haslanger, who has failed to demarcate these vary different ends for which concepts can be improved.

Thus, MCE is distinct from ameliorative analysis, instead forming a distinct part of analysis of this type. Both MCE and NCE are examples of Haslanger's ameliorative analysis, just with differing purposes.

3.2 Naturalist conceptual engineering

The second type of conceptual engineering we distinguish is *naturalist conceptual engineering* (NCE). This is conceptual engineering undertaken with scientific goals in mind. The method will not differ from MCE, but the goals and desiderata for the concept will rely on scientific norms rather than moral ones – such as explanatory power, measurability, or concordance with our best scientific understanding of the world (for some detailed examples of such epistemic goals, see Carballo 2020). One example of a concept that has undergone engineering within biology has been the concept of species, where a diversity of concepts have been proposed, each with particular benefits for their role in different sciences (Mayr 1992). Other examples of concepts that are engineered primarily for a scientific role could include genes, species, models, measurement, etc. There is an incredible diversity of different roles concepts play in science, and we do not dare to begin listing all of them here. Rather, we offer the following pluralist account of NCE:

> Naturalist Conceptual Engineering = (i) The assessment of concepts, categories, and classificatory systems according to scientific norms, (ii) determination of their relevant context and purposes to which they are and should be put to use, (iii) reflections on and proposal for how to improve them, and (iv) proposals for and active participation in the implementation of the suggested improvements.

It might be natural to take NCE to be a version of Carnap's (1950) concept of 'explication'. It would be a mistake, however, to equate all conceptual engineering within science as explication or to think that only Carnapian explication is a justified form of conceptual engineering within science. It is not our goal here to ameliorate Carnapian explication, a particular method with a rather clear but limited role in science, but instead we shall offer a brief survey of the diverse ways NCE can occur in science. Importantly, unlike Carnapian explication, NCE is instead a set of methods of which Carnapian explication is a mere member.[7] In Carnap's own words:

> The task of making more exact a vague or not quite exact concept used in everyday life or in an earlier stage of scientific or logical development, or rather of replacing it by a newly constructed, more exact concept, belongs among the most important tasks of logical analysis and logical construction.

[7] Novaes (2018) argues that Carnapian explication while not explicitly about moral or political values, is implicitly endorsing Enlightenment values such as emancipation and freedom – in line with Carnap's political stance (see Carus 2007). Ordinarily, however, Carnapian explication is merely seen as scientific concept refinement, which is the received view we shall follow. If Carnapian explication is political, then it would be even more of a mistake to equate it with NCE, rather than a hybrid of the two kinds of conceptual engineering discussed here.

> We call this the task of explicating, or of giving an explication for, the earlier concept. (Carnap 1947: 8–9)

Carnap, like other members of the Vienna Circle, is mostly concerned with the usefulness of concepts in the formulation of scientific laws. In this vein, he discusses the taxonomic concept of 'Pisces' as a scientific explication of the folk concept of fish – unlike its vague and intuitive counterpart within folk terminology, it is better able to play a role in scientific laws (1950). We can take a term common in ordinary language and try to refine it in various ways to study a phenomenon in nature, which then in turn leads to a further refinement of our terms. This is roughly what Carnap has in mind when he speaks of explication as a procedural improvement of our concepts. The philosophy of science, however, has long moved on from such a narrow conception of scientific progress and Carnapian explication is better conceived as one way among many possible ways of engaging in NCE.

The primary purpose of distinguishing NCE from MCE is to bring some philosophical clarity into debates about concepts that play a role in both science and society where the differing goals we put these concepts to use are kept obscure. Concepts are intended for a diversity of roles, and purposes are manifold. From Haslanger's discussion of race and gender to old philosophical debates on consciousness and welfare, many of the philosophical discussions attempt to untangle a muddled field of concepts, categorizations, and classificatory schemes. Where there is confusion such as this, philosophy has a useful role to play. To do so, however, we need to disentangle the different roles, functions, and purposes for which respective concepts are put to use.

Here the context in which the concept is used is key. This may be disappointing to those who try to provide a unified picture of all of conceptual engineering, but such monist aspirations should be resisted. Indeed, philosophers can still play useful roles, but they need to dive into the actual conceptual debates, taking constant care to resist the temptation of extrapolating from one conceptual debate to all others. We have given some examples of the uses of MCE and NCE for different concepts, however there are also many cases in which we will not have a single role for our concept and instead will want both. Let us now move on to illustrate how the competing demands of these two kinds of conceptual engineering can come into conflict in practice, through the examples of two concepts which squarely fall into both the scientific and moral/political domain: human health, and animal welfare.

4. Case studies: Health and welfare

4.1 Health

The first example of a concept that has received a significant amount of philosophical attention is health, or to expand it a little more: health and pathology. Indeed, the conceptual debate on the status of these

concepts is at the core of the philosophy of medicine. Yet, decades of debates have seemingly moved us farther away from a consensus, rather than towards it. Rather than taking this expansion of views as a mere indication of more philosophers entering the debate, we can see it as a conflict between different demands for which the concepts of health, disease, and pathology are put to use.

That this option has received fairly little attention is due to the widespread acceptance of conceptual analysis as the only tool needed to settle the debate (Schwartz 2007; Lemoine 2013; Schwartz 2014). This has perhaps been the result of an intention to eventually arrive at something like a list of necessary and sufficient conditions that would help us to demarcate 'normal' from 'pathological' states. But the search for something like a conceptual essence or correct criteria of application may have been overly naive. To think about health without the social and biological context in which these concepts are used is bound to lead to widely differing accounts. It is not surprising that the field is usually described as a conflict between so-called 'naturalists' who try to make these notions into legitimate scientific notions, and 'normativists' who emphasize non-epistemic values (especially moral ones) that go into our judgements of health and disease.

Instead of framing the debate as one with two competing camps fighting over the 'one true' definition of these concepts it may be more useful to think about the different goals these camps are interested in. For this purpose, our distinction between MCE and NCE provides a useful tool to think about the conflicts between the purposes to which we put these terms. We may even come to realize that there is no single concept that can play the different roles sufficiently well, due to inevitable trade-offs. Yet, it is only by paying attention to these larger trade-offs between naturalist and normativist goals that we can shed light on what our concepts *ought* to mean.

The Canadian philosopher of science Ian Hacking (1991) has previously made a similar argument when he pointed out that our concept of 'child abuse', which was once precisely defined and operationalized as 'battered child syndrome', has changed much over the years to encompass an ever-greater number of different actions deemed vile and in turn reformed "our values and our moral codes" about what is and what is not appropriate treatment of children (p. 253).[8] This is tempting, of course, because we can use moral language to advance our moral values, but in doing so the concept has also lost its grasp on something once deemed to be a natural phenomenon that we could make scientific generalizations about. There can be numerous trade-offs between designing a concept for the purposes of capturing a phenomenon in nature and for the purposes of morality that we need to pay attention to. Our point here is not to argue that this is how things *must* be. That would be a legitimate topic for a further stand-alone paper. Health

[8] We thank Paul Griffiths for alerting us to this early recognition of trade-offs.

merely provides a beautiful case for a concept in which MCE and NCE may pull in vastly different directions. As the Welsh psychiatrist Robert Evan Kendell once put it:

> The most fundamental issue, and also the most contentious one, is whether disease and illness are normative concepts based on value judgments, or whether they are value free scientific terms; in other words, whether they are biomedical terms or sociopolitical ones. (Kendell 1986: 25)

In thinking about the goals of 'health' and 'disease' it appears strikingly hard to maintain a strict binary dichotomy in which health is either entirely value-free or entirely political. Yet, such positions have been defended at length. A more useful approach may be to accept that the concept has both naturalist and ethical components to it. It is not every *undesired* or *unvalued* state, but neither is it just some biologically *dysfunctional* arrangement of a body. In thinking about the concept in terms of conceptual engineering, applying the distinction we introduce here, it may then be useful to accept that there could be different weightings we place on each dimension. How important are biological facts in thinking about disease? How should we think about unjust social arrangements that may be the true cause of a bodily difference being considered 'wrong'? In his earlier work, Christopher Boorse (1975) advocated for a distinction between two kinds of health, one of which is purely naturalist and opposed to *disease*, and one that ethical or normative and opposed to *illness*. Others, like Jerome Wakefield (1992), have argued for the need for a hybrid account that combines a dysfunction criterion with a notion of harm. How should we evaluate such competing proposals?

We suggest to put at the heart of the debate the question of what roles these concepts ought to play. As Kukla (2014) once put it: "in considering the best definition of health, we need to keep clearly in view the theoretical and practical purposes to which we want to put the concept, while keeping an open mind as to how unified a definition is possible" (p. 516). One radical solution such an approach may reveal is that these various uses of the concept cannot adequately be covered in a single concept. We may then become pluralists, or even try to eliminate some of its dimensions altogether if there are other concepts that could better play the required roles. Instead of thinking about whether a bodily state deserves medical treatment, we may stop to ask whether it is a pathology, and even ask the (perhaps more interesting) question of whether the condition decreases wellbeing or autonomy. This may expand the goals of medicine, but that is of course at least an option. The goals of medicine can be changed just like those of any institution or enterprise.

In the case of health and disease we can legitimately embrace some skepticism that the widely different purposes for which the concepts are put to use among different groups can be satisfied with a single concept. If a purely naturalist concept of disease is misused to discriminate

against gay or transgender people because of their alleged dysfunction and reduction of fitness, it may not be good enough to insist on a strict is-ought distinction or that no moral and political conclusions will follow from a purely biological set of criteria alone. There are good reasons to not label these conditions as pathological, precisely because that can inevitably be misused to justify homophobia or transphobia. Here, we are engaged in MCE. Yet, how to respond to the potential charge that 'ideology' is changing our concepts, by those who maintain that they are 'just stating biological facts?' Here is where we can emphasise the distinction between NCE and MCE, and the differing goals and desiderata accompanying each. In thinking about concepts, we have to keep in mind the sociopolitical context in which these concepts are put to use. To make a concept 'tidier' for the purpose of philosophical simplicity under NCE, would be a terrible mistake if it creates real harms or unjust treatment of already vulnerable groups. Too much thinking about health and disease in the philosophy of medicine has happened from the ivory tower. This is why Haslanger's call to pay attention to these issues has been so important and transformative for philosophy as a whole.

Normativists differ widely in their motivations. This is perhaps not surprising since concepts can play myriad roles, that can be good in some cases, while being bad in others. It may be inevitable to need to take tradeoffs between different goals seriously. Neither MCE nor NCE must point to a single solution. The concept of pathology, for instance, may play different theoretical roles in evolutionary biology, cancer research, veterinary science, animal production science, immunology, and economics. Reducing them to a single concept may not be possible. Fortunately, meta-philosophical discussions are becoming ever more prevalent in the philosophy of medicine. We hope that our distinction between MCE and NCE will help to progress its core debate further. There could hardly be a more illustrative example for the usefulness of it. Having discussed a concept in which a possible separation is likely, we will now turn to a concept in which MCE and NCE work much more closely together.

4.2 Animal welfare

Another example of a concept that lends itself to both MCE and NCE is animal welfare. This is a concept which plays both moral and scientific roles that have historically grounded deliberations about which concept to use. It is also one for which there is still much current debate about which concept to employ, often without making explicit the particular values or desiderata in play. This thus provides a fruitful example of the use of this distinction within conceptual engineering. Animal welfare is a useful case study for the application of these methods, as it is both the subject of scientific study and of moral deliberation. It is therefore important to ensure we have a concept that fulfils both roles.

Both MCE and NCE are relevant and important in deliberations as to the best concept of animal welfare.

Animal welfare plays a scientific role within animal welfare science. Scientists study welfare both to gain increased understanding of the behaviour and biology of the animals, and to provide information that can inform moral and policy decision-making. Scientific measurement of welfare will rely heavily on the concept of welfare in use, determining which measurements are considered valid, as well as which conditions will turn out to impact welfare. This tension can be seen in cases where different concepts are employed, leading to different conclusions. For example, there has been ongoing disagreement on the permissibility of the use of sow stalls between those holding different welfare concepts; with the different concepts employed leading one group to endorse and another to reject their use (Croney and Millman 2007). Both groups take themselves to be measuring welfare, but have identified a different target and are thus talking past each other. To implement NCE on the concept of animal welfare, we must identify the desiderata for this concept within the scientific role, and then assess which concepts best fill the criteria. Two such desiderata are measurability and fundamentality. As a target for scientific investigation, it is important for welfare to be something that scientists can measure. Additionally, if we take welfare to be the appropriate relevant target for investigation within welfare science, it must also be something which is fundamental. That is, whatever fills the role must be something which is not itself simply a property of or proxy for some other state, something itself an intrinsic part of the characterisation of welfare (Browning and Veit 2024).

Animal welfare also plays an important moral role. It is a central concept within much of animal ethics and is typically considered to have moral importance. Animal welfare is thus relevant to decisions made by legislators, producers and consumers with regards to housing and treatment of animals. Expenditure of time and resources on animal welfare improvements requires a clear understanding of what welfare is, in order to ensure interventions are effective. Otherwise, the risk is that efforts may be wasted on providing conditions that may appear to increase welfare without actually doing so. Similarly to NCE, applying MCE to animal welfare requires identification of the desiderata for this concept within the moral/social role from which to assess the suitability of different concepts for meeting these criteria. For the moral role, the most important criterion seems to be that the concept tracks something we take to be of normative significance. We also want it to be capable of identifying those bearers of moral worth. Welfare being morally important, then those individuals capable of experiencing welfare should therefore form part of the moral community. Establishing a welfare concept will make rulings on which individuals fall within this group, and our assessment of which they rule in and out will affect our judgment as to the suitability of the concept.

We thus have two different roles for our welfare concept, with differing values and constraints. The next step is to assess the candidate concepts according to the desiderata. Which set of desiderata we use will depend on the context of assessment – whether we are considering the scientific or the moral role for welfare and thus whether we are undertaking NCE or MCE. Although it is possible to advocate pluralism, so that the concept used varies depending on the different application (Veit and Browning 2021), in many cases the close link between the outputs of animal welfare science and moral deliberations means that it will be important to use the same concept for both so that our concepts in both areas to refer to the same entity. There are four primary concepts of animal welfare in use today: subjective welfare, physical welfare, teleological welfare and preference-based welfare (Browning 2020). Although this assessment could itself be the subject of a full paper, here we will briefly describe each of these concepts and indicate their suitability according to the desiderata, to give an idea of how this would be applied. As discussed, for the scientific role, the requirements are for a concept that is measurable, and fundamental; while for the moral role, the requirements are for a concept that delineates something of moral value and can identify its bearers.

The physical concept of animal welfare was common in earlier versions of animal welfare science, taking animal welfare to consist in some set of physical functionings – bodily health and comfort. These concepts were selected almost entirely as a result of NCE rather than MCE. Physical states are easy to measure but it is more difficult to make a case as to why they should matter morally. Most of the reasons we have for thinking that poor physical functioning matters is due to its negative effects on the subjective experience of animals, which means it is not fundamental. It also fails to delineate the bearers of moral worth from those without – if physical functioning is what matters, then animals, plants and microorganisms may all be said to have a welfare equally worth considering, and this is not a view many wish to accept.

Natural living, or teleological accounts (Browning 2019) of welfare, emphasise the overall 'flourishing' of an animal according to its nature; generally focussing on the performance of natural behaviours. This concept is popular within the general public (Lassen et al. 2006; Vanhonacker et al. 2008), as on the surface it appears to do well for both NCE and MCE - being based in species biology as well as seemingly morally important. However, a deeper examination shows this concept does poorly on both NCE and MCE. It is not easily measurable, as identifying what count as natural behaviours and how much of their performance counts as good welfare is unclear (Veasey et al. 1996). Additionally, it is not obviously of moral importance – although intuitive to some, there is no strong account of how or why naturalness should matter morally (Browning 2019). It also fails to delineate the boundar-

ies of our moral circle in line with common intuitions – again, all organisms are capable of natural functioning, but it is not common to extend equal moral consideration this distance.

The subjective welfare concept is perhaps most common in current use, and describes welfare as the balance of experience of positive and negative mental states (affects). It meets the requirements for MCE in being normatively significant, as the capacity for subjectively experienced pleasure and suffering provides cause for moral concern and delineates the boundaries for moral consideration. However, it is often taken as unsuitable for a scientific role (Dawkins 2017) due to inaccessibility to measurement; a view that is less common where it is accepted that subjective experience is not epiphenomenal - having no causal impact on the world - but instead that there must be behavioural and physiological effects of mental states, which we can then detect and can form the basis for measurement.

Another popular welfare concept is a preference-based account of welfare (Dawkins 2003), which takes welfare to consist in satisfaction of preferences. This concept also does well for MCE: preference-based accounts of wellbeing are common (e.g. Griffin 1986) and most consider the satisfaction of their desires to be highly valuable. The use of this concept has been advocated in welfare science (Dawkins 2017) due to one of the demands of NCE – they are easily measurable through preference-based behavioural tests, in which animals are presented with different options and observed to see which they choose, and how hard they will work to attain it. The primary deciding factor between these two is whether preferences or experience is more fundamental - whether subjective experiences are valuable because they are desired, or whether desires are valuable because they create positive experiences.

The example of animal welfare perfectly demonstrates the distinction between NCE and MCE. As shown for the different welfare concepts, these two roles can be in conflict, and the needs of NCE and MCE may not both be met within a single concept. However, when we are able to establish a concept that meets the requirements of both, this gives stronger reasons to adopt that concept, particularly when we want the two roles to intersect. This example has focussed primarily on the first three components of conceptual engineering - assessment of the concepts, determination of relevant context/purposes and proposals for improvement.

The stage of proposal for implementation of improvements is an important one - perhaps the hardest (Chalmers 2020) – but will be highly contextdependent and require more discussion than we can provide here. However, the discussion of teleological welfare can provide some insight – the discrepancy discussed between public and expert views on welfare also gives reason to think that part of the role of adopting or modifying scientific concepts is to assist in educating the general public as to the preferred concept and the reasons underlying its choice.

In this case, informing the public as to the defects within this concept could have wide-reaching effects in the decisions made by consumers and advocates for animal welfare. Switching out a teleological concept for an alternative, such as subjective welfare, will alter which conditions might be thought important for welfare. For example, zoo visitors often prefer seeing monkeys in naturalistic island-type enclosures, but in actuality, cage-style exhibits often provide more climbing surfaces and opportunities for activity, promoting good subjective welfare (Browning and Maple 2019).

Currently, the subjective or preference concepts appear to do best under both types of conceptual engineering and would thus be preferred in the contexts for which we want the two to coincide. More generally, this serves as an illustration of the method of applying NCE and MCE, in identifying the desired role(s) for a concept and the different desiderata required for each. In this case, it may be possible for a concept to fill both roles and we will prefer one that does. Making this process explicit can help shine light on previously muddled debates about which concepts should be preferred, demonstrating the context-sensitivity and the need to be clear about the goals for use of the concept.

5. Conclusion

To conclude, our distinction between *naturalist conceptual engineering* (NCE) and *moral conceptual engineering* (MCE) refers to the differing goals of conceptual engineering, and their associated desiderata, rather than its methods. While Carnapian explication is traditionally associated with the formal methods and tools of logic and the natural sciences, amelioration is often understood as the qualitative improvement of concepts by drawing on the humanities and social sciences (see Novaes 2018). Our pluralist account of conceptual engineering combines these and other forms of 'concept improvement' as mere methodologies for a diverse set of practices that fall under the umbrella term 'conceptual engineering'. Some of these ameliorative methods can be used for both scientific desiderata and moral/political purposes – something we may very well consider a feature, rather than a bug.

Nevertheless, where philosophers become engaged in highly divisive debates about concepts with very little consensus, it should be our task to alleviate and disentangle muddled conceptual confusions. In this paper we have illustrated how this might be done, using the examples of health and animal welfare. Care must be taken not to extrapolate from one conceptual debate to all others. To do so, however, we need to separate the functions, goals, and purposes for which particular concepts are put to use. After all, concepts, categories, and classificatory systems play too many roles as to allow for a single simple monist account of 'conceptual engineering'. This will require us to move much closer to examination of scientific practice, history, and sociology – and hence

endorse a pluralist and pragmatic form of conceptual engineering that is much closer in spirit to Neurath than it is to Carnap.

Rather than endorse a particular methodology for the revision of our concepts, we recognize a diversity of way concepts can be changed and altered, which can in turn be grouped into two general categories with different goals. They are distinguished by their functions, not their methods. This can be seen as a foundation for future metaphilosophical research on these two kinds of conceptual engineering and the potential use of this distinction in untangling the philosophical and scientific debates on controversial concepts such as race, gender, disability, and mental disorder.

References

Boorse, C. 1975. "On the Distinction between Disease and Illness." *Philosophy & Public Affairs* 5 (1): 49–68.

Brigandt, I., and E. Rosario. 2020. "Strategic Conceptual Engineering for Epistemic and Social Aims." In A. Burgess, H. Cappelen, and D. Plunkett (eds.). *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, 100–124.

Browning, H. 2019. "The Natural Behavior Debate: Two Conceptions of Animal Welfare." *Journal of Applied Animal Welfare Science* 22 (4): 325–337.

Browning, H. 2020. *If I Could Talk to the Animals: Measuring Subjective Animal Welfare*. PhD diss., Australian National University.

Browning, H., and T. L. Maple. 2019. "Developing a Metric of Usable Space for Zoo Exhibits." *Frontiers in Psychology* 10: 791.

Browning, H. and W. Veit. 2024. "Animal welfare science, performance metrics, and proxy failure." *Behavioral and Brain Sciences* 47: E70.

Burgess, A., H. Cappelen, and D. Plunkett (eds.). 2020. *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Cappelen, H., T. Gendler, and J. P. Hawthorne (eds.). 2016. *The Oxford Handbook of Philosophical Methodology*. Oxford: Oxford University Press.

Cappelen, H., and D. Plunkett. 2020. "A Guided Tour of Conceptual Engineering and Conceptual Ethics." In A. Burgess, H. Cappelen, and D. Plunkett (eds.). *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, 1–27.

Carballo, A. P. 2020. "Conceptual Evaluation: Epistemic." In A. Burgess, H. Cappelen, and D. Plunkett (eds.). *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press, 304–332.

Carnap, R. 1947. *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Cartwright, N., J. Cat, L. Fleck, and T. E. Uebel. 2008. *Otto Neurath: Philosophy between Science and Politics*. Cambridge: Cambridge University Press.

Carus, A. W. 2007. *Carnap and Twentieth-Century Thought: Explication as Enlightenment*. Cambridge: Cambridge University Press.

Chalmers, D. J. 2011. "Verbal Disputes." *Philosophical Review* 120 (4): 515–566.

Chalmers, D. J. 2020. "What Is Conceptual Engineering and What Should It Be?" *Inquiry* 63 (9–10): 954–970.

Croney, C., and S. Millman. 2007. "Board-Invited Review: The Ethical and Behavioral Bases for Farm Animal Welfare Legislation." *Journal of Animal Science* 85 (2): 556–565.

Dar-Nimrod, I., and S. J. Heine. 2011. "Genetic Essentialism: On the Deceptive Determinism of DNA." *Psychological Bulletin* 137 (5): 800–818.

Dawkins, M. S. 2003. "Behaviour as a Tool in the Assessment of Animal Welfare." *Zoology* 106 (4): 383–387.

Dawkins, M. S. 2017. "Animal Welfare with and without Consciousness." *Journal of Zoology* 301 (1): 1–10.

Devitt, M. 1981. *Designation*. New York: Columbia University Press.

Godfrey-Smith, P. 2003. *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago: University of Chicago Press.

Griffin, J. 1986. *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford: Clarendon Press.

Hacking, I. 1991. "The Making and Molding of Child Abuse." *Critical Inquiry* 17 (2): 253–288.

Haslanger, S. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Noûs* 34 (1): 31–55.

Haslanger, S. 2005. "What Are We Talking About? The Semantics and Politics of Social Kinds." *Hypatia* 20 (4): 10–26.

Haslanger, S., and J. Saul. 2006. "Philosophical Analysis and Social Kinds." *Proceedings of the Aristotelian Society, Supplementary Volume* 80: 89–143.

Jackson, B. B. 2014. "Verbal Disputes and Substantiveness." *Erkenntnis* 79 (1): 31–54.

Jenkins, C. S. 2014. "Merely Verbal Disputes." *Erkenntnis* 79 (1): 11–30.

Kendell, R. E. 1986. "What Are Mental Disorders?" In A. M. Freedman, R. Brotman, I. Silverman, and D. Hutson (eds.). *Issues in Psychiatric Classification: Science, Practice and Social Policy*. New York: Human Sciences Press, 23–45.

Kornblith, H. 2002. *Knowledge and Its Place in Nature*. Oxford: Oxford University Press.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kukla, R. 2014. "Medicalization, 'Normal Function,' and the Definition of Health." In J. Arras, E. Fenton, and R. Kukla (eds.). *The Routledge Companion to Bioethics*. New York: Routledge, 539–554.

Lassen, J., P. Sandøe, and B. Forkman. 2006. "Happy Pigs Are Dirty! – Conflicting Perspectives on Animal Welfare." *Livestock Science* 103 (3): 221–230.

Lemoine, M. 2013. "Defining Disease beyond Conceptual Analysis: An Analysis of Conceptual Analysis in Philosophy of Medicine." *Theoretical Medicine and Bioethics* 34 (4): 309–325.

Linquist, S., E. Machery, P. E. Griffiths, and K. Stotz. 2011. "Exploring the Folkbiological Conception of Human Nature." *Philosophical Transactions of the Royal Society B* 366 (1563): 444–453.

Machery, E. 2016. "Experimental Philosophy of Science." In J. Sytsma and W. Buckwalter (eds.). *A Companion to Experimental Philosophy*. Malden: Wiley-Blackwell, 475–490.

Machery, E., P. Griffiths, S. Linquist, and K. Stotz. 2019. "Scientists' Concepts of Innateness: Evolution or Attraction?" In R. S. D. Wilkenfeld (ed.). *Advances in Experimental Philosophy of Science*. London: Bloomsbury, 172–201.

Margolis, E., and S. Laurence (eds.). 1999. *Concepts: Core Readings*. Cambridge: MIT Press.

Mayr, E. 1992. "Species Concepts and Their Application." In M. Ereshefsky (ed.). *The Units of Evolution: Essays on the Nature of Species*. Cambridge: MIT Press, 15–26.

Nersessian, N. J. 1989. "Conceptual Change in Science and in Science Education." *Synthese* 80 (1): 163–183.

Neurath, O. 1913. "Probleme der Kriegswirtschaftslehre." *Zeitschrift für die gesamte Staatswissenschaft* 69 (3): 438–501.

Neurath, O. 1921. *Anti-Spengler*. Munich: Callwey.

Novaes, C. D. 2018. "Carnapian Explication and Ameliorative Analysis: A Systematic Comparison." *Synthese* 195 (3): 1013–1034.

Papineau, D. 2013. "The Poverty of Conceptual Analysis." In M. Haug (ed.). *Philosophical Methodology: The Armchair or the Laboratory?* London: Routledge, 166–194.

Pollock, J. 2019. "Conceptual Engineering and Semantic Deference." *Studia Philosophica Estonica* 12: 81–98.

Quine, W. V. O. 1960. *Word and Object*. Cambridge: MIT Press.

Schwartz, P. H. 2007. "Decision and Discovery in Defining 'Disease.'" In H. Kincaid and J. McKitrick (eds.). *Establishing Medical Reality: Essays in the Metaphysics and Epistemology of Biomedical Science*. Dordrecht: Springer, 47–64.

Schwartz, P. H. 2014. "Reframing the Disease Debate and Defending the Biostatistical Theory." *Journal of Medicine and Philosophy* 39 (6): 572–589.

Sytsma, J., and W. Buckwalter (eds.). 2016. *A Companion to Experimental Philosophy*. Malden: Wiley-Blackwell.

Thagard, P. 1992. *Conceptual Revolutions*. Princeton: Princeton University Press.

Thomasson, A. L. 2017. "Metaphysics and Conceptual Negotiation." *Philosophical Issues* 27 (1): 364–382.

Uebel, T. 2019. "Vienna Circle." In E. N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). Stanford: Metaphysics Research Lab, Stanford University.

Vanhonacker, F., W. Verbeke, E. Van Poucke, and F. A. Tuyttens. 2008. "Do Citizens and Farmers Interpret the Concept of Farm Animal Welfare Differently?" *Livestock Science* 116 (1–3): 126–136.

Veasey, J. S., N. Waran, and R. Young. 1996. "On Comparing the Behaviour of Zoo-Housed Animals with Wild Conspecifics as a Welfare Indi-

cator, Using the Giraffe (*Giraffa camelopardalis*) as a Model." *Animal Welfare* 5: 139–153.

Veit, W. 2020. "Model Pluralism." *Philosophy of the Social Sciences* 50 (2): 91–114.

Veit, W., and H. Browning. 2021. "Perspectival Pluralism for Animal Welfare." *European Journal for Philosophy of Science* 11 (1): 1–14.

Wakefield, J. C. 1992. "The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values." *American Psychologist* 47 (3): 373–388.

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.5
Received: December 9, 2025
Accepted: March 10, 2026

Deliberation, Action and Freedom

DAVOR PEĆNJAK
Institute of Philosophy, Zagreb, Croatia

In this article, I try to present a cumulative argument for libertarianism concerning free will and mainly from a theistic perspective. First, I present and develop further an argument from Pećnjak (2018) that if we have "actish phenomenal feeling" (Ginet 1990) that in a certain situation we can genuinely decide between action A and action B, and that God is not a deceiver, then we have a good reason to believe in libertarianism. I connect this line of reasoning with St Anselm's view on freedom of the will, namely that we have genuinely open possibilities and that we can persevere in what is good, on our own, and that this perseverance is a choice we make from ourselves. In the last part, I present certain experimental evidence (Schulze-Craft et al. 2016) that agents can voluntarily stop an action which started as an unconscious brain process. A certain congeniality of these three ways gives us firm ground to believe in libertarianism.

Keywords: God; freedom of the will; deliberation; libertarianism; determinism; St Anselm of Canterbury.

1. Introduction

In this article, I shall examine a few important notions in the free will debate and I shall argue for the libertarian position from a theistic perspective.[1] The free will debate has many sides and it is not possible to embrace all of them in one article, but it is in order to explicate just some fundamentals before we go into specifics here. When I say that I shall take a theistic stance, it means that I take for granted that the universe and human beings are created by God, and that I shall use, in argumentation, certain facts about God's attribute of benevolence. I shall not discuss the problem of God's foreknowledge and human free-

[1] This work has been fully supported by the Croatian Science Foundation under the project "Intentionality and Modes of Existence" [IMEX IP-2022-10-5915].

dom here, because it is in fact a problem different in nature from the problem I discuss here.

2. Libertarianism and determinism

Libertarian position, in general, is an incompatibilist one which means that libertarians think, first, that concepts of "determinism" and "freedom" ("freedom of the will", "freedom of action") are in no way reconciliable – if there is any kind of freedom, then determinism could not be the case, and this is a libertarian position. Those who are also incompatibilists, but hold that determinism is the case, think that there is no kind of freedom and so are hard determinists. It is clear that compatibilist think that determinism and freedom are compatible so, that even if we embrace determinism, there could be a reconciliatory notion of freedom, and so freedom could exist even in a deterministic world. I shall leave a discussion of compatibilism, and I shall say only that I think that neither version of compatibilism is in the slightest a tenable theory.

Determinism can be defined in various ways. One of the standard definitions is that if we take any instant of the universe and have a complete description of it (all the positions of particles, entities, their velocities, momentum, charge etc), together with the laws of that universe, all other states in the development of the universe are uniquely determined – only one history of the universe is possible.

It means, if determinism is true, that we can take any instant we would like, and all other instants uniquely follow, no matter whether they are past or future relative to the instant we have chosen. But, usually, we take determinism in the way that proceeds from the past to the future – we may take the very first instant of the universe and, if determinism is the case, then each and every future instant of the universe uniquely follows – it is completely determined what they will be. The unfolding of the universe is nomologically necessary if determinism is the case.

We can interpret what is said as a logical thesis – an instant with the laws *entail all other states and instants in a unique way; we can interpret it as a causal thesis - an instant of the universe, with the laws causes* all other states and instants in a unique way. More precisely, if causal determinism holds, then, according to laws of nature that obtain, each event is causally necessitated by a previous event. I shall rely more on determinism as a logical thesis here, but nothing special depends on it.

There are several versions of libertarianism. The main versions are event-causal libertarianism, agent-causal libertarianism and non-causal libertarianism. I shall not argue here for any of these specific versions, but for libertarianism in general. It will be enough, for present purposes, that it can be shown that under the same conditions, a

subject or an agent can do otherwise than in fact he did, and that we can have persuasive reasons which show that an agent is in control over what he does, and that what he does is not necessitated in any way, especially not by the factors on which an agent does not have a control.

In recent articles (Pećnjak 2018, Pećnjak and Anić, forthcoming), it is argued that if determinism is the case, then there is no deliberation, and that this result is unacceptable; so it tells against determinism. It is also argued (Pećnjak 2018) that if we have what Ginet (1990) calls *actish phenomenal feeling* and God exists, then we have free will in a libertarian sense. I do not provide specific or additional arguments for God's existence, as I did not provide them in a previous text. I assume the existence of God. If this is not enough for some of the readers, I plead here, as I pleaded in Pećnjak (2018), that the reader insert here her or his favourite argument(s) for the existence of God.

I shall first give a summary of these two arguments (for more details, see Pećnjak 2018) and then, on the basis of them, we may proceed to the further main points of this article, which will consist of connecting these previous arguments with the free will theory of St Anselm, and certain recent empirical research about free will.

3. Thoughts about deliberation

Deliberation is a process through which we come to a solution on what to do in a certain situation when we are faced with, at least seemingly, various different possible outcomes. This process is a complex process and can extend in time and may be interrupted. However, we shall treat it here as a one continuous process for the ease of exposition because I think that nothing crucially depends on treating it as such in order to explicate philosophically relevant matters concerning deliberation in connection with free will. When we deliberate, we examine various beliefs we have, various desires, inclinations, reasons, values we hold, we mentally simulate various outcomes and their implications, we imagine, simulate and evaluate possible future situations. We try to follow the basic laws of logic in these processes, where applicable, and we try to care about the contents of those mental states we process (and according to them, to infer further steps in deliberation).

So, we may say that deliberation is a complex mental process involving various mental items – so it is a species of a more broadly conceived category of action. Deliberation is action – mental action. Furthermore, as an action, it is an event. Again, deliberation typically consists, in fact, of many mental events that are bound together under the same agenda, so we may say that deliberation is a complex mental action, i.e., a complex mental event. It seems that the subject who deliberates is in full control of this complex mental event of his.

But now, suppose that determinism is the case. What becomes of deliberation? If determinism is the case, then it overarches the so-called

process of deliberation as well (Pećnjak and Anić, forthcoming). But this overarching is then so disastrous that nothing of our deliberation will remain. How is it so? Since deliberation is a process that develops in time, each step of this mental process is then fully determined. It is fully and uniquely determined by some initial state and the laws of nature. We may also say that every step in our thinking would then be necessitated by previous states and laws of nature – it would be nomologically necessitated.

Any mental process that happens in a deterministic world is itself determined, so it is not possible, nomologically, for any mental process not to occur in such a world; moreover, it is not possible, nomologically, for any part or any step of that process to be different from what it is. The subject which has this process does not have even the possibility to start or to refrain from this mental process at time t when this process in fact starts (nor at any moment t before or after). So, nothing in this mental process is under the control of the subject. It's the other way around: the process controls the subject. Since the process is a consequence of the initial state of the universe and the laws which obtain there, no one can do, even in thinking, otherwise than he in fact did.

Action that would stem from this kind of mental process would not be a free action because it would stem from the (mental) process which is itself fully determined and, secondly, the action itself, (as well as the mental process from which it stems) is also fully and uniquely determined (already) by the initial state of the universe and the laws which obtain there. So I conclude, there would be no deliberation in a deterministic world, nor would there be agents in the usual meaning of the concept "agent". Why?

Because we can explain what seems to be deliberation – that mental process which occurs for the subjects in a deterministic worlds - by the factors which are evidently and fully outside the control of the subject who "deliberates"; each and every step in the process is a unique consequence or is entailed by factors which are beyond the grasp and control of the subject. So, "deliberation" and its sequence and content are in "control" of the initial state and the laws of nature and not of the subject – the subject is, in fact, only a passive observer of what happens to him. Likewise for his physical activities ("actions") that stem from these mental processes: – the actions which occur by subjects in a deterministic world are due only to the factors which are evidently and fully outside the control of the subject who "acts"; each and every step in the process is a unique consequence or is entailed by factors which are beyond the grasp and control of the subject.

But this result is unacceptable, I think, for our world.

It seems that we really do deliberate and it seems that we are really agents. It seems that we really have open possibilities in front of us – both in deliberating and in action. It seems that we can do otherwise than what we in fact, at the end, did. How and why is this possible?

4. God is not a deceiver and actish phenomenal feeling

I shall use some thoughts from Desartes (1911) and Ginet (1990) to show this (see also Pećnjak 2018 for earlier statements about this part of the argument). First, it is not hard to establish that God is not a deceiver. Here, I shall be very brief. God is traditionally conceived as a morally perfect being. Of course, God has many other attributes, but these other attributes are not pertaining for our needs here. Being who is morally perfect would not systematically mislead the beings which are His creations (and seemingly have consciousness, capacity for morality, rationality and a capacity to lead a diverse life and advance and improve themselves). So, a morally perfect being would not deceive – so, God is not a deceiver.

The concept of "actish phenomenal feeling" is introduced by Carl Ginet (1990) and it refers primarily to our intentional mental actions. But it also refers to the will when it comes to making something with our body or its parts. It is something that accompanies our intentional saying something to ourselves or, for example, when we consciously try to make occurent some mental item from our memory. But let Ginet speak for himself: "The act of mentally saying peu is a different sort of mental event from the unbidden occurrence of that word in one's mind. ... The unbidden occurrence is not an act. And, most importantly, the mental act does *not* consist of an event just like an unbidden occurrence *plus* its having a certain extrinsic relation to the subject. Rather, the mental act differs from the passive mental occurrence *intrinsically*. The mental act has what we may call (for lack of a better term) an *actish* phenomenal quality. This is an extremely familiar quality, recognizable in all mental actions, whether it be mental saying, mental forming an image, or willing to exert force with a part of one's body. ... This quality is intrinsic to and inseparable from the occurrence of the word in my mind when I mentally say it." (Ginet 1990: 13) So, even when we try consciously and intentionally to make a certain (physical) action with our body or its parts (moving hands or legs in order to do something, or even just moving them), we have this actish phenomenal feeling which is such that tells us that we and nothing else is the ultimate origin of making that action.

So, it applies to both mental actions and mental precursors of our physical actions, which stem from our mentally formed will. If this is true, and I think it is, then we can be in control of our thoughts, i.e., beside sometimes "unbidden" occurrences in our consciousness of various contents – be it words, phrases, beliefs and whatever other content that may become conscious – we can and very often are, in control of what we think, when we think and to what conclusions we came. So, we can really have a process properly called deliberation, because it can be itself freely done in each of its steps and the action which eventually follows, also is free. Action is in fact free even in a twofold sense. It is a proceeding of a freely obtained result of deliberation (namely of will

and intention thus formed), and even when we start to do an action, we still have a possibility to refrain from it.

If we are the ultimate origins of making mental or physical action, it is up to us what we make, and from these it follows that we could have done otherwise than we did, both mentally and physically.

5. St Anselm's and Anselmian approach

Now, to complete the theistic argument. Since God created human beings and since human beings have this "actish phenomenal feeling", and since God is not a deceiver, then we are not deceived by Him that we have freedom of deliberation, and, hence freedom of the will and freedom of action. We really have these freedoms. Namely, God would not let us have something (namely that "actish phenomenal feeling"), that would lead us so often, massively, on an everyday basis into a completely false stance about our freedom in deliberating, will and action.

It seems to me that with what I offer here is congenial with St Anselm's theory of free will. St Anselm formulated his theory of free will in the following treatises: "On Free Will" and "Why God Became Man" (St Anselm of Canterbury 1998). Let me explicate just the basics of his theory with some adaptations for modern use. He considers that human beings, as created rational beings, have the power to choose between mutually exclusive options. How is that so? First of all, traditionally, God is conceived as the supreme creator, which means that each and every created being is dependent on God; everything that exists is dependent on God. According to St Anselm, God creates a situation for human beings that consists of at least two mutually exclusive possibilities for future action. Human beings, according to such a creation of the situation by God, have desires to pursue both of them. St Anselm adds that this situation, which consists of a twofold desire system for pursuing incompatible courses of action, is necessitated by God. If God created human beings with only one desire to pursue and act in only one way, that pursuing and action would not be free and it would be fully necessitated by God, so by factors wholly outside the control of human beings as human agents. If something is fully outside the control of a supposed agent and if the agent could not do anything about it, then determinism would be the case. So, if human beings were to have only one desire or affection to do just one action, then in fact that desiring and action would be determined by God, and in fact human beings would be only God's deterministic puppets. But, according to St Anselm, it is not so. Having two different desires to act in incompatible ways leaves room for genuine opting for one of them. Everything that is created and everything that is different from God depends on God. Still, God has given the power to choose to his crown of creation, to human beings. This power, as a power, and that we have it, really depends on God, but the power itself is such that nothing is in advance determined by it or

by God. Human beings themselves use this power to choose between possibilities. Nothing and noone else, including God Himself, determines the results of what is in the scope of this God-given power which human beings have. But, possibilities and awareness of these possibilities are something that is given by God. In especially morally relevant situations of choice, human beings know what is right to do, but they are also aware of doings which are not right. Both of these, as possibilities and awareness of them, are God-given. But, according to St Anselm, God has given to everyone to know what is right, and we can persevere in the righteousness of will – namely, to will, and then to do, what is right. In St Anselm's words, we can persevere in truth and righteousness. This very perseverance, as mental action, is not something that is determined, even it is not determined by God. It is up to each human being to persevere – we have the power to do it, this is what is God-given only. So, no one and nothing, besides each human being, chooses for him/herself; then, as agents, each one of us determines for himself or herself what to will and what to do according to this will. We may choose what is not right, but we do not thereby lose our power for freedom of the will, as it is a power; we only do not then "persevere in what is right." The power is always there as is the possibility to persevere in what is right. Choosing sin or a morally bad action, we thereby do not lose this power of will (to persevere in righteousness). Though, adds St Anselm, the most free is the being who never stops persevering in what is right. This is so because every time we choose what is right, we choose it freely and just for the sake of righteousness itself. Perseverance in good is freely done because the subject who wills it, wills it on his or her own; nothing necessitates perseverance, though there is a possibility that the will succumbs to something that is not good, but nothing necessitates this as well.

All that is said, though St Anselm is concerned with moral situations, is also applicable to choosing in non-moral contexts. Subject, or an agent, is aware of two or more possibilities what to will and how to act, and it is up to him or her to choose one possibility and then bring it to action. Noone else and nothing else determines the choosing and a decision, which is forming the will, what to do. Using a bit of St Anselm's language, a human being can persevere in holding his or her decision, as the will to do something, and can persevere in doing actions on the basis of the persevered will, and nothing necessitates this persevering, as well as nothing necessitates refraining from this persevering.[2] Freedom is given at a moment of choice when one course of action has been chosen, which happens between two equal desires for pursuing an action in incompatible ways. So, the required indeterminacy for freedom of the will and freedom of the action (which follows)

[2] For careful and detailed modern interpretation of St Anselm's theory of freedom and free will, see Rogers (2008, 2015). See also Gwozdz (2009) and Nash-Marshall (2008).

is located in a moment of choice when one course of action has been chosen rather than another. This is a moment in which God does not intervene, and agents choose completely by themselves for pursuing that one course of action. So, this is called choosing *a se*. Rogers (2015) rightly calls st Anselm's theory of free will agent-causal theory[3]. There is no special power to choose between incompatible courses of action; the agent him/herself just perseveres in one course of action, choosing it and performing it (Rogers 2015 calls it per-willing). This is an event because it happens in time (Rogers 2015: 97). So, how does this fare with the traditional claim that God is the creator of everything that exists? Anselmian answer would be, as Rogers points out (2015: 97), that "God is the cause of all that exists, but He is not the cause of all that happens." Certain events are caused by beings to which God has given the ability to make choices and perform actions based on those choices. So, sovereignity of God is not at all jeopardized by giving agents libertarian agent-causal freedom of the will and action.[4]

Now, we can return briefly to the first theistic argument put forward in the first part of the article to see its full strength, combining it with Anselmian view of freedom of the will. There, we said that, according to Ginet, we have an actish phenomenal feeling when we deliberate and decide – this actish phenomenal feeling gives us that we feel that at the moment of choice we by ourselves directly just make a decision, which has an intentional structure, and that nothing else determines this (and not even God). We make a decision just by making it by ourselves, by that very mental act. This actish phenomenal feeling is located within deciding between two incompatible desires that God has given or created in us, according to St Anselm, where there is genuine indeterminacy at the moment of choice and that indeterminacy is God given that we by ourselves can fully make that decision and pursue one rather than the other course of action and so persevering or per-willing in one chosen course of events. As we said, God is not a deceiver, and

[3] For agent-causal theories in contemporary philosophy, see O'Connor (2000), Chisholm (1964), Clarke (1993, 2003: chapters 8-10), Griffith (2007). For a different neo-scholastic libertarian theory of free will, so-called "Dual Sources", see Grant (2019).

[4] In his interesting study, Peter Furlong (2019) examines what the consequences of divine determinism would be. He defines divine determinism borrowing Heath White's proposal, which consists of two claims: a) The facts about God's will wholly determine every other contingent fact, and b) The facts about God's will will explain every other contingent fact (Furlong 2019: 15). Given our arguments in this text, using the Anselmian approach to free will, it is not so. It is not so because the facts about God's will do not wholly determine every other contingent fact, nor is every other contingent fact explained by the facts about God's will. Since human beings can choose by themselves which course of action to pursue, these courses of action (which are contingent facts) are not determined by God's will, nor are they explainable by God's will (apart from the fact that God *allows* them). So, there is no threat of divine determinism. According to the view presented here, God determines "everything that exists, but does not determine everything that happens" (Rogers 2015: 97).

because this actish phenomenal feeling is pretty much straightforward, we have strong reason to believe that it is right and true in presenting to us that we just by ourselves directly bringing, on our own, a decision and action that follows.

6. Experimental evidence

With the things said till now, certain experimental results go hand in hand. I do think that good philosophy does not need experimental data to confirm it, but nevertheless, I find these data also to signal that in this world, for which we are the most interested, we have genuine incompatibilistic freedom of the will and action. Though those philosophers, such as Mark Balaguer (2010), who think that in the end the freewill problem boils down to the empirical question, would be delighted. Neurological electroencephalogram experiments of Schultze-Kraft et al. (2016) showed that there is a possibility for subjects to stop their intended action until 200 ms before the beginning of the physical execution of the action. So-called spontaneous movements are preceded by the onset of brain activity (and it is called readiness potential). This onset of neural brain activity is not already conscious when it does begin and it is unconscious for some time on. It seems that sometimes certain brain neural activity can be even four seconds long (Soon et al. 2013.) before subjects tell that they are conscious of their choosing, as only then they feel it as a conscious will and conscious intention to do something. So, it may seem that determinism reigns because if our actions are products of non-conscious neural brain activity, then human beings are not free – something else determines what human beings do – namely, laws of nature and previous states on which these laws operate. That would threaten what we said above about the possibility of having freedom of will, deliberation and action. When we became conscious of our will, it seems that it is too late – it is only an illusion that we decide and will something on our own consciously, though all this would be a product of other factors – unconscious brain events – over which we do not have an influence. But is it so? Further experiments show that it is not.

"As early as a second before a simple voluntary movement, a so-called readiness potential is observed over motor-related brain regions" (Schultze-Kraft et al. 2016: 1080), which is not conscious.[5] The question that is highly important is then, "...whether a person can still exert a veto by inhibiting the movement after the onset of the readiness potential"? (Schultze-Kraft et al. 2016: 1080). Experimental results of Schultze-Kraft et al. (2016), show that persons can act in such a way that they can issue a veto "even after onset of this preparatory brain process" (Schultze-Kraft et al. 2016: 1080), which is non-conscious.

[5] Though these researchers also warn that what is the exact causal role of these signals is in fact not definitely solved.

Their results "suggest that humans can still cancel or veto a movement even after the onset of the readiness potential. This is possible until the point of no return, around 200 ms before movement onset. However, even after the onset of the movement, it is possible to alter and cancel the movement as it unfolds." (Schultze-Kraft et al. 2016: 1084). This suggests that human beings can consciously and voluntarily stop an action that started unfolding as an unconscious neurological process. So it seems that we are not at the mercy of our unconscious processes. We can voluntarily change the course of our own actions.

7. Conclusion

Several different points point to the same conclusion, so we may regard them as a kind of cumulative evidence for libertarianism: we see that there is a plausible way to construct human will and human action as having the property of libertarian free will. It seems that actions, both mental and physical, are up to us in the sense that we could have done otherwise than what, in fact, we did.[6]

References

St Anselm of Canterbury. 1998. *The Major Works*. Edited by Brian Davies and G. R. Evans. Oxford: Oxford University Press.

Balaguer, M. 2010. *Free Will as an Open Scientific Problem*. Cambridge: MIT Press.

Chisholm, R. 1964. *Human Freedom and the Self*. Lindley Lecture. University of Kansas.

Clarke, R. 1993. "Toward a Credible Agent-Causal Theory of Free Will." *Noûs* 27: 191–203.

Clarke, R. 2003. *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.

Descartes, R. 1911. *The Philosophical Works of Descartes*. Translated by E. S. Haldane and G. R. T. Ross. Cambridge: Cambridge University Press.

Furlong, P. 2019. *The Challenges of Divine Determinism*. Cambridge: Cambridge University Press.

Ginet, C. 1990. *On Action*. Cambridge: Cambridge University Press.

Grant, W. M. 2019. *Free Will and God's Universal Causality*. London: Bloomsbury Academic.

Gwozdz, T. 2009. "Anselm's Theory of Freedom." *The Saint Anselm Journal* 7 (1): 1–13.

Griffith, M. 2007. "Freedom and Trying: Understanding Agent-Causal Exertions." *Acta Analytica* 22 (42): 16–28.

Nash-Marshall, S. 2008. "Free Will, Evil, and Saint Anselm." *The Saint Anselm Journal* 5 (2): 1–23.

Pećnjak, D. 2018. "Free Deliberation." In F. Grgić and D. Pećnjak (eds.). *Free Will and Agency*. Berlin: Springer.

[6] I would like to thank reviewers for their helpful comments.

Pećnjak, D. and Anić, Z. Forthcoming. "A Note on Determinism and Deliberation."

O'Connor, T. 2000. *Persons and Causes*. Oxford: Oxford University Press.

Rogers, K. 2008. *St Anselm on Freedom*. Oxford: Oxford University Press.

Rogers, K. 2015. *Freedom and Self-Creation: Anselmian Libertarianism*. Oxford: Oxford University Press.

Schultze-Kraft, M., D. Birman, M. Rusconi, C. Allefeld, K. Görgen, S. Dähne, B. Blankertz, and J. D. Haynes. 2016. "The point of no return in vetoing self-initiated movements." *Proceedings of the National Academy of Science of the USA* 113 (4): 1080–1085.

Soon, C. S., A. Hanxi He, S. Bode, and J. D. Haynes. 2013. "Predicting free choices for abstract intentions." *Proceedings of the National Academy of Science of the USA* 110 (15): 6217–6222.

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.6
Received: April 24, 2024
Accepted: May 17, 2025

Temporal Integration and the Basis of Moral Equality

TIMOTHY J. NULTY
University of Massachusetts Dartmouth, North Dartmouth, USA

Belief in universal moral equality—that all people have equal moral status—has wide cultural and political acceptance and holds favor with many philosophers. I argue against status-parity by offering a novel account of the temporal integration of persons. Some persons have much more robust temporally extended selves which are partially constituted by a special class of future-directed interests. In addition, empirical research indicates persons vary significantly in the degree of affective connection to their futures selves. While some persons have a cognitively complex and affectively strong connection to their future selves, other persons' future selves are analogous to strangers. Differences in temporal integration directly affect recognized morally relevant properties such as rationality, autonomy and moral agency and entail differing degrees of personhood. Crucially, some differences in degrees of personhood will also involve differences in kind. The paper then critically engages with attempts to provide a basis for moral equality.

Keywords: Moral equality; temporal integration; degrees of personhood; future-directed interests.

1. Preface and initial characterization

Morally relevant properties such as self-awareness, autonomy and rationality which afford persons a higher moral status than merely sentient beings vary considerably among persons. The challenge for advocates of moral status parity is to reconcile this variability with the claim that all persons have equal moral status. Providing such grounding, as anticipated by Nietzsche's 1882 "Parable of the Madman," requires finding some way to support the idea that all persons have equal moral worth in the absence of a justifying metaphysical story. Nietzsche writes, "This tremendous event is still on its way, still wander-

ing… This deed is still more distant from them than the most distant stars—*and yet they have done it themselves*" (Nietzsche 1974: 181–182). Although we ourselves had killed god according to Nietzsche, the implications of that event have *still* not reached our ears. We retain certain moral values which now lack any clear metaphysical underpinning.[1]

Peter Singer has taken the death of god seriously. His discussions of animal suffering and the value of different types of lives are based on secular interest utilitarianism. Despite Singer's self-proclaimed attempt to construct a moral theory free from religious influence, it is a matter of dispute how successful he has been. Brian Leiter notes, "Parfit and Singer think of themselves as vanguards in this movement, a claim rich in irony for any student of Nietzsche" (Leiter 2019: 386). Leiter explains that Singer's belief one ought to treat all equal interests equally is the kind of suspect egalitarianism Nietzsche challenged. Nonetheless, Singer's utilitarianism leads to some very inegalitarian results.

Singer claims species membership does not afford *homo sapiens* special moral status. However, the lives of *persons* do have special moral status compared to the lives of non-persons. When it comes to killing, what matters, according to Singer, is not species membership but whether the being to be killed is a person: "No objective assessment can support the view that it is always worse to kill members of our species who are not persons than members of other species who are" (Singer 1993: 117). It is worse, in general, to kill persons than non-persons precisely because of the kinds of interests involved. Persons have future-directed interests while non-persons lack such interests: "For preference utilitarians, taking the life of a person will normally be worse than taking the life of some other being, since persons are highly future-oriented in their preferences…In contrast, beings who cannot see themselves as entities with a future cannot have any preferences about their own future existence" (Singer 1993: 95). While being a person is a threshold condition for having a life with greater moral significance, Singer does not provide any argument that it is equally wrong to kill persons.

Since some persons have more future-directed interests than others, killing one person can violate more future-directed interests than killing someone else. Given Singer's framework, it can be worse to kill one person rather than another when more interests are violated. Admittedly, the concept of moral status has a derivative role—if it has any role at all—in consequentialist theories.[2] Singer equates the wrongness of killing with the degree of harm caused by the killing, and harm itself is defined in terms of violated interests. Singer abandons any robust

[1] See Husi (2017: 388): "If people were equally beloved by a supreme God … we might have a vindication for status-parity without worrying about equalizing grounds."

[2] Christiano (2015: 61) and Husi (2017: 385) both make a similar point.

foundational notion of moral status grounded in factual equality when he writes: "The plain fact is that humans differ, and the differences apply to so many characteristics that the search for a factual basis on which to erect the principle of equality seems hopeless" (Singer 1993: 18). Rather than talk of beings (e.g., humans or persons) having equal moral status, Singer adopts the normative principle that we should give equal consideration to equal interests.

Egalitarians find this result unacceptable because they endeavor to show that all persons have certain rights equally, such as the right to life. Regardless of the degree of harm death causes, many egalitarians would argue it is equally wrong to kill persons because they have equal moral status or are worthy of equal respect.[3] While some egalitarians have assumed the equal moral status of persons, others have recognized the need to ground status equality in morally relevant properties or through "relation-first" and "practice-based" approaches to equality.[4]

I offer a novel account of temporal integration based on an empirically-informed examination of future-directed interests. I argue the way interests constitute persons threatens foundational property-based accounts of moral status. An elaboration of the concept of future-directed interests is essential to capturing the ontological distinctness of persons. This elaboration will follow two tracks. The first is a philosophical analysis of variations in the content of future-directed interests. I distinguish between self-shaping future-directed interests (SSFIs) and non-self-shaping future-directed interests (NSFIs). I argue SSFIs have much greater weight than NSFIs. SSFIs have greater weight because they constitute our identity as individuals and involve a uniquely personal kind of autonomy.

The second track, relying on empirical research, shows the affective intensity with which persons have various interests about their future selves, or whether they have certain future-directed interests at all, varies with the perceived degree of continuity they have with their future selves. This perceived degree of temporal connectedness affects the rationality of choices persons make about their futures. The person-constituting role of some future-directed interests, especially when informed by empirical research, suggests that the inegalitarian results of Singer's consequentialist approach generalize and undermine attempts at grounding status parity on morally relevant properties.

Future-directed interests are intimately connected to the rationality, autonomy, self-awareness and temporal unity of persons. These commonly recognized morally relevant properties must be understood in relation to future-directed interest as holistically constituting the be-

[3] See McMahan (2002: 235–245) for one example, especially his rejection of the Time-Relative Interest account.

[4] See for example McMahan 2002; Carter 2011, 2018; Christiano 2015; Miklosi 2022 and Lipshitz 2024. Sangiovanni 2015 offers a relation-first approach which is critiqued by Floris 2019, 2020. See Rozebloom 2018 for a practice-based account of moral equality.

ing of persons. These properties are morally relevant precisely because of the kind of being they constitute. A proper ontological understanding of persons justifies awarding them a higher moral status than non-persons in part because persons are a *different order of being*. Because there are degrees of personhood, it also provides strong reasons against the claim that all persons have equal moral status. Crucially, some differences in degrees of personhood will also be differences in *kind*.

2. Self-shaping future-directed interests (SSFIs)

Among the properties associated with personhood, awareness of a temporally extended self is arguably the most central property. Non-human primates and very young humans lack the robust sense of self typical of adult humans. Developmental psychologists maintain the sense of self emerges over time and is enhanced by language acquisition. A toddler's understanding of itself as a temporal entity is much more limited than an adult's understanding. Beings are not considered persons if they lack a sense of self. A person's life transcends the biological and becomes autobiographical.

The philosophical tradition recognizes the importance of self-awareness. Locke claims a person "… can consider itself as itself, the same thinking thing, in different times and places" (Locke 1979). Singer (1993) writes approvingly of Joseph Fletcher's criteria for personhood which include self-awareness, a sense of past and a sense of future. DeGrazia lists "self-awareness over time" and "complex forms of consciousness" in addition to other essential properties such as "autonomy, moral agency, rationality and capacity for intentional action" (DeGrazia 2007: 319-320). Self-awareness is necessary for autonomy and moral agency. Rationality is understood partially in terms of self-interest, and self-interest requires having self-awareness. We have good grounds for believing a temporally extended self—the organism's awareness of its specific identity over time—is a necessary condition for personhood and *grounds* others features such as rationality, autonomy and moral agency.

Singer writes, "Very often, it [killing] will make nonsense of everything that the victim has been trying to do in the past days, months, or even years" (Singer 1993: 95). Singer provides examples of the kinds of interests he has in mind: "For example, a professor of philosophy may hope to write a book demonstrating the objective nature of ethics; a student may look forward to graduating; a child may want to go for a ride in an aeroplane" (Singer 1993: 90). The concept of future-directed interests is left largely unanalyzed by Singer. In the examples Singer provides, some future-directed interests appear to be intimately connected to one's current and future identity as a person (e.g., a professor finishing a book or a student graduating college), while other future-directed interests are unlikely constitutive of our current and future identity (e.g., the child wanting a plane ride). We can, therefore, distin-

guish between self-shaping future-directed interests (SSFIs) and non-self-shaping future-directed interests (NSFIs).

There are three types of future-directed interests: (1) interests with content that refers to the future, but the content is non-self-referential or non-autobiographical; (2) interests with content that refers to the future and which either explicitly requires the person's continued existence or entails it; and (3) interests with content that refers to the future, entails the person's continued existence, and is about becoming a particular kind of self. The interests in this last category have content through which a person aims at actualizing some potentiality of their current self. I will refer to them as "self-shaping future-directed interests" or SSFIs.

Type-1 future-directed interests refer to the future, but their content does not require or entail the person's future existence. I could desire that my great grandchildren have good lives, but that future-directed interest does not involve *my own* continued existence. Killing a person does not violate type-1 future-directed interests since the person's continued existence is not necessary for satisfying the interest.

Type-2 interests require the person's continued existence. A person might desire to retire to a nice tropical location, and that future-directed interest does entail the person's continued existence. Thus, killing a person will violate type-2 future-directed interests. Singer's (1993) discussion of the chimp Figan exemplifies type-2 future-directed interests. Figan, noticing a banana, does not want the dominant chimp Goliath to take it. Figan intentionally avoids looking at or trying to obtain the banana until Goliath has left the area. Satisfying that type-2 future-directed interest requires Figan's continued existence. Young children have many type-2 interests such as Singer's example of the child wanting to ride in a plane.

Type-3 future-directed interests introduce an element unlikely to be found in very young children and non-human primates. These interests are connected to our self-creation by selecting possibilities and attempting to actualize them. Singer's examples of a professor wanting to finish writing a book and a college student wanting to graduate are type-3 interests. These future-directed interests not only require the possessor's continued existence, but they involve possessors taking a stand on their own identity. The person I am now depends, in part, on the kinds of activities into which I project myself. Type-3 future-directed interests play a special role in the temporal unification of the person because they allow the person to gather up large temporal expanses of their lives. Type-3 future-directed interests allow persons to exercise a uniquely personal form of autonomy: self-creation. This class of future-directed interests is also essential to moral reasoning and being a moral personality. Determining what sort of person someone wants to be in a moral sense, and contemplating the consequences of a possible action for the agent's own moral standing, involve type-3 future-directed interests.

3. Personal identity

Personal identity theories provide criteria for the diachronic identity of persons. They do not tell us *to what degree* those beings are persons. The same criteria may apply to both borderline persons and paradigmatic persons, but how those criteria are realized will differ significantly. Moreover, there are reasons for thinking the diachronic identity of a person is distinct from the temporal unity of that person's self. Having a self is a property shared by all persons, but the degree to which persons have selves varies, and this will affect the degree to which something is a *particular person*.

There are at least three aspects of persons that admit of degrees: (1) following Parfit, there are degrees to which person at t1 is the same person at t n+1; (2) there are degrees to which something is a person at all (e.g., toddlers and some non-human primates); and (3) the degree to which something is a particular person. Person A is more of a particular person than person B when person A has a richer and more complex set of individuating psychological states. For example, Figan who is less *generally* person-ish than a typical human adult is also less of a *particular* person because Figan has fewer and less complex psychological states that constitute his particular identity. He has a less robust sense of self. Thus, the degree of general personhood limits the degree to which something is a particular person.

Psychological theories of personal identity distinguish between *psychological connectedness* and *psychological continuity*. Consider Parfit's description of the Psychological Criterion:

> (1) There is psychological continuity if and only if there are overlapping chains of strong connectedness. X today is one and the same person as Y at some point in the past if and only if (2) X is psychologically continuous with Y, (3) this continuity has the right kind of cause, and (4) it has not taken a branching form. (5) Personal identity over time just consists in the holding of facts like (2) to (4). (Parfit 1984: 207)

Humans can have many direct psychological connections without being a person to the degree of a typical adult. The development from non-person to person is a gradual one. When a sense of self emerges between 2.5 and 3.5 years of age, toddlers do not have the same complex sense of self as adults. It is likely non-human primates who might be persons also lack the complex sense of self of adult humans. Yet, toddlers and non-human primates have enough of the right kinds of psychological states and capacities to be considered borderline persons.

The psychological connections that comprise the continuity of toddlers and non-human primates will include type-2 future-directed interests (e.g., Figan wanting the banana or the child wanting a plane ride). The psychological connections of some paradigmatic persons will include not only type-2 future-directed interests, but the more complex and self-aware type-3 future-directed interests. Both the toddler and Figan have a diachronic identity as minimal persons, while persons

with elaborate type-3 interests will be *individuals* to a greater degree. This distinction is similar to Peter Unger's concepts of "core psychology" and "distinctive psychology". Distinctive psychological states are those which are not shared by all persons and which are instead unique to us as individuals or only shared with a few other persons (Unger 1990). This interest-based distinction is also similar to Christian Perring's (1997) distinction between general personhood and particular personhood.

Think of Relation R more broadly in terms of content and the connections between the content of individual mental states. There are varying degrees of unification or what we might call "integration." What makes a psychological connection over time strong? Parfit states, "Connectedness can hold to any degree… Since connectedness is a matter of degree, we cannot plausibly define precisely what counts as enough" (Parfit 1984: 206). Parfit explains degrees of connectedness in terms of the *number* of direct connections: "But we can claim that there is enough connectedness if the number of connections, over any day, is at least half the number that hold, over every day, in the lives of nearly every actual person" (Parfit 1984: 206). The connections are treated as psychological events independent of their specific content. We have reasons to modify Parfit's account of psychological connectedness. There are forms of temporal connectedness which are distinct from the number of overlapping psychological connections. Some persons will have richer and more complex forms of psychological continuity compared to other persons even when the number of overlapping connections is the same, and these persons will have selves with greater diachronic unity.

We can consider Kierkegaard's view of the self to help us understand how persons' selves can vary in their degree of temporal unity. Kierkegaard's aesthete might have many hedonistic future-directed interests in his attempt to avoid boredom from moment to moment. The aesthete might have *more* direct psychological connections than the person striving to live an ethical or religious life. For Kierkegaard, what gives unity to the person is the nature of what one pursues or that into which persons project themselves. Temporal unity increases as a person progresses from the aesthetic life, to an ethical life, and ultimately a religious life. Kierkegaard describes what is "most inward and holy" in humans as the "unifying power of the personality" (Kierkegaard 2005: 13). Kierkegaard maintains that the unity of the individual is an achievement closely connected with the way one lives:

> The choice itself is decisive for the content of the personality, through the choice the personality immerses itself in the thing chosen, and when it does not choose it withers away in consumption… That which has to be chosen stands in the deepest relationship to the chooser…. (Kierkegaard 2005: 13)

We can develop Kierkegaard's insight as an addition to Parfit's numerical account of strong connectedness. There are at least two important factors that temporally unify the self: (1) the extent to which the

future-directed interests relate to the person's identity as a particular individual, and (2) the extent to which the contents of the person's future-directed interests are related to each other.[5] The professor's interest to complete a book is intimately connected with his identity and his other interests to be promoted, to be invited to speaking engagements, etc. Thus, the professor's future-directed interest to complete the book exemplifies a strong degree of connectedness because it is integrated and self-referential. The college student desiring to graduate, much like the professor, has future-directed interests which are self-shaping. The college student's interest to graduate is closely connected to many other current and future-directed interests such as passing exams and earning high course grades. In both cases, there is also the unification of large temporal expanses. The child who desires a plane ride scores much lower, as would someone desiring to play video games to avoid boredom. These isolated interests are not linked in any significant way to other interests. They are not strongly connected to the person's identity as a temporally extended self, and the temporal range is much shorter. Much like borderline persons only strive to *get* things and not to *become* something, some persons have impoverished future-directed interests which primarily aim at acquiring things rather than developing into a particular self.

The degree of unity over time is not solely a matter of the number of direct psychological connections as Parfit argues; rather, the unity over time which constitutes a person is a product of the content of various psychological states including future-directed interests. Some people have more or greater SSFIs than others, so they have selves with greater unity over time and instantiate a greater degree of personhood. SSFIs have objectively greater weight than NSFIs because they are what constitute us as specific persons. Preventing the professor from finishing the book or the student from graduating would be a greater wrong than preventing the child's plane ride or preventing someone from playing video games. The interest to develop oneself into a particular person and to shape one's life accordingly is the highest expressions of personhood.

This result is consistent with the intuition shared by some people that the death of a person pursuing a promising future is more of a tragedy than the death of someone indifferent to his future self, e.g., a hedonist only interested in the next pleasurable experience through drugs, food or sex to avoid boredom. Similarly, the death of Figan the chimp is less of a tragedy than the death of a student pursuing a college degree. Compare the death of Figan to historical discrimination against women and minorities. Figan's death thwarts all his type-2 interests. Discrimination against women and minorities prevents the satisfac-

[5] See McMahan (2002: 75). McMahan identifies three factors which comprise psychological unity, but importantly, McMahan does not include mental content about one's own identity as relevant to unity over time.

tion of many type-3 future-directed interests. Women and minorities were prevented historically from creating themselves as they saw fit. Slavery would be one of the most compelling examples. The inhibition of self-creation is a greater violation of autonomy than the inhibition of type-2 interests. The intuition that thwarting type-3 interests is worse than thwarting type-2 is explained by recognizing that type-3 interests involve a different *kind* of autonomy than type-2 interests, not merely a difference in degree, and the former has greater worth than the latter.

4. The empirical research

Psychologists and cognitive scientists are generating a growing body of research exploring our relationships to our future selves and the real-world implications of varying degrees of connectedness. The research shows connectedness to our future selves plays a role in our health, financial decisions, ethical behavior, academic success, and relationships. Empirical investigations about connectedness to our future selves offer additional ways of conceptualizing degrees of temporal unification.

Empirical research focuses on three factors which contribute to our degree of temporal connectedness: similarity, vividness and positivity. Hershfield (2011) reviews research showing these three factors affect the number and intensity of connections to persons' future selves:

> Critically, then, the degree to which an individual feels disconnected from his or her future self should correlate with the degree to which that individual discounts future rewards. The more continuity a person shares with his future self—that is, the more that future self feels like a direct extension of who he is now—the more motivated he will be to act in ways that will benefit himself in the future. Conversely, the more the future self feels like a stranger—that is, the more disconnected a person is from his future self—the less motivated he will be to plan for the future. (Hershfield 2011: 34)

Psychometric tasks have been developed to assess *future self-similarity*. These tasks determine the degree of similarity persons feel to their future selves. Differences in degree of future self-similarity have a variety of behavioral and attitudinal consequences. There is a correlation between degree of similarity and temporal discounting. For example, persons who perceive their future selves as less similar will discount the value of saving for retirement: "In line with our prediction, we found a significant positive correlation between perceived similarity to the future self and the number of assets that had been accumulated over time" (Hershfield 2011: 36).

Some individuals relate to their future selves as strangers. Phenomenological differences in degree of connectedness are exhibited at a neurological level. For some individuals, thinking of their future selves elicits neural activity identical to thinking about strangers while for others such future-directed thoughts elicit neural activity similar to thinking about their current selves. These neurological differences—

whether the brain treats the future self as more like a stranger or more like the current self—affect temporal discounting:

> As expected, there was individual variability in these neural differences: for some participants, thinking about the future self elicited neural activation patterns that were almost exactly like patterns that were associated with thinking about another person; for other participants, thinking about the future self elicited neural activation patterns that were more or less in line with patterns associated with thinking about the current self. In line with our prediction, participants who showed the biggest difference between activation elicited by the current self and activation elicited by the future self also discounted future rewards most steeply. (Hershfield 2011: 37)

Individuals who view their future selves more like strangers are generally less interested in preparing for the future, and when they do have such interests those interests are less important. Such persons are less likely to make choices today to benefit themselves in the future. A student, for example, will be less inclined to study to increase his future chances of success and might opt instead to spend time having fun.

How vividly the future self is perceived is the second factor which determines the degree of connectedness. When individuals have vivid and realistic representations of their future selves, they are more likely to save for the future. In one experiment, students are shown a virtual reality version of their future selves while another group of students is shown a virtual reality version of their current selves. "As hypothesized, those participants who were exposed to their future selves were subsequently more likely to allocate money toward a hypothetical retirement savings account than were control participants" (Hershfield 2011: 38). Vividness is also an important factor in predicting delinquency. According to Jean-Louis van Gelder, et al (2013), individuals unable to contemplate their future self with a sufficient degree of vividness are more likely to engage in delinquent behavior:

> The tendency to live in the here and now, and the failure to think through the delayed consequences of behavior, is one of the strongest individual-level correlates of delinquency. We tested the hypothesis that this correlation results from a limited ability to imagine one's self in the future, which leads to opting for immediate gratification. Strengthening the vividness of the future self should therefore reduce involvement in delinquency. We tested and found support for this hypothesis in two studies. (van Gelder 2013: 974)

Positivity is the third factor affecting the degree of connectedness to the future self. Individuals who feel more positive about their future selves also show a greater degree of connection to those future selves. As with the previous two capacities, positivity correlates with rational behavior which benefits one's future self.

Recall morally relevant properties such as temporal self-awareness, rationality, autonomy and moral personality. The capacity to have vivid future self-representations about which one can feel a high degree of similarity and positivity is clearly a very rich form of self-awareness.

The empirical research shows the greater the degree of similarity, vividness and positivity, the more individuals exhibit rational self-interest and autonomy about the future. Parfit characterizes rational self-interest in the following way: "A rational agent should both have, and ultimately be governed by, a *temporally neutral* [my emphasis] bias in his own favor. It is irrational for anyone to do what he believes will be worse for himself" (Parfit 1984: 307). Given this account of rational self-interest, individuals who score higher in similarity, vividness and positivity are more rational than those who exhibit those capacities to a lower degree.

When the future self is perceived as a stranger the person has less interest now in making decisions to benefit the future self. When the future self is perceived as a direct extension of the current self, the person has a much stronger interest in benefiting the future self. Similarly, persons positively inclined toward their future selves have greater interests in benefiting their future selves. Consider DeGrazia noting the chicken's lack of "temporal self-awareness" (DeGrazia 2007: 317) or Singer's discussion of persons having "preferences about their own future existence" (Singer 1993: 95). The chicken and other non-persons lack any conception of a future self. They don't have a temporally extended self about which they could be aware. Borderline persons may have some minimal, short-term sense of a future self, but such a self is conceptually thin compared to most paradigmatic persons. Based on empirical research, some persons have an impoverished sense of their own future existence. In DeGrazia's terms, they have a limited form of "temporal self-awareness" compared to those who perceive a high degree of similarity with their future selves. For some persons, the concept of *my own* future existence is much more robust. Those persons who perceive their future selves as strangers cannot conceptualize their future as belonging to *them* in the same way those who see their future selves as a direct continuation of their current selves.

5. Against equal status

We've seen the temporal unity of persons varies because of the kinds of future-directed interests (e.g, type-2 vs. type-3) and the degree of affective connections to their future selves. Moreover, the kind of future-directed interests and the degree of affective connectedness bear significantly on degrees of autonomy, rationality and moral personality, all of which are morally relevant properties. These considerations provide a morally relevant basis for a distinction between higher-type and lower-type persons. In *Thus Spoke Zarathustra*, Nietzsche writes, "What is the ape to man? A laughing-stock or painful embarrassment. And just so shall man be to the Superman" (Nietzsche 1961: 42). We find similar themes in the same work where man is described as "…a rope, tied between animal and Superman…" (Nietzsche 1961: 43). In *Beyond Good and Evil*, Nietzsche claims, "In man *creature* and *created*

are united…" (Nietzsche 1966: 154). We can develop these Nietzschean insights in a morally relevant way.

Non-persons cannot endeavor to create themselves. They are merely created beings because they lack type-3 future-directed interests. For Nietzsche, human beings are created animals and have some capacity for self-creation, but that capacity and its actualization vary significantly. Based on my earlier analysis, there are two kinds of higher-type persons: (1) those persons that have a higher degree of general personhood, and (2) those persons who are *individuals* to a greater degree (i.e., more of a particular person). Among persons who are equal in terms of general personhood, some have mental states and behavioral patterns shared by many, say largely dictated by social media rather than by autonomous, rational choice, while others possess greater numbers of genuinely differentiating psychological traits.[6] This latter group of persons will be *individuals* to a higher degree. Among persons with a higher degree of general personhood, some persons will be self-creators while others will be passively created by cultural factors.

Being a true individual involves more than having distinctive psychological traits; it requires self-creation through genuinely autonomous choices. Self-creators exhibit a different and higher kind of autonomy and rationality than non-self-creators. The autonomy of authentic self-creation is a different kind of autonomy than Figan's behavioral autonomy or the average person's choice of television show. Autonomous self-creation requires intellectual autonomy which is again different in kind than the simpler kinds of behavioral autonomy. Thus, the higher degree of particular personhood instantiated by self-creators involves differences in *kind*.[7]

Nietzsche stressed the integrative quality of higher-type individuals and its connection to rationality. Rationality allows persons to control their impulses, unify them in self-creation, and affords persons the freedom of self-mastery (i.e., greater autonomy). Kaufmann explains this point: "Reason is the 'highest' manifestation of the will to power…. because these skills enable it to develop foresight and to give consideration to all the impulses, to organize their chaos, to integrate them into a harmony…." (Kaufmann 2013: 230). Nietzsche's notion of integration parallels the results of section 2 and section 3 of this paper. The highest type persons are self-creators who are richly integrated, complex, and strongly temporally unified via the kinds of future-directed interests

[6] Consider Perring: "Gradually increasing psychological properties might separate an individual in one species from individuals in other species (general personhood) while doing little to separate her from other particular persons. The development of a distinct personality does increase the particular personhood of an individual" (Perring 1997: 183).

[7] The point here is relevant to Sangiovanni's claim: "our rights against being treated as an inferior—and hence to equal moral status in my terms—vary along with our capacities to develop and maintain an integral sense of self" (2014: 104). On my view, self-creators have a significantly greater capacity to develop and maintain an integral sense of self.

they have and the interrelation among those interests. Nietzsche refers to these persons as "sovereign individuals" (Nietzsche 1967: 59).

The account of temporal integration offered here is richer than Carter's (2018) adaption of Parfit's account. Carter focuses on the temporal continuity of volitional capacities related to human agency. Carter, like Parfit, correctly believes diachronic integration is scalar, and he maintains the number of connections varies as well as the temporal range (e.g., short-term or long-term). Following his earlier (2011) work, Carter argues we should ignore variations in diachronic integration as a matter of opacity respect "…by refraining from taking into account her degree of diachronic integration beyond the minimum threshold…" (Carter 2018: 835). Carter's argument for equality despite variations in diachronic integration depends on both his use of range properties and opacity respect. I return to both ideas shortly.

Christiano (2015) presents the grounding of moral status problem as a trilemma: "One, the status of persons is grounded in the extent to which they have certain distinctive traits. Two, persons have the status conferring traits to relevantly different degrees. Three, persons have equal status" (Christiano 2015: 55). Egalitarians respond to this apparently inconsistent triad by arguing that either claim one or claim two is false in the hope of preserving claim three.

One response to the trilemma claims the threshold of rationality matters, but changes above the threshold do not matter because "… a change from below the threshold to above the threshold involves some kind of substantial transformation of the nature of the being involved while changes above the threshold do not involve such substantial transformation" (Christiano 2015: 73). Christiano offers the example of giving a chimpanzee the level of rationality typical of most humans and claims this would "transform the being into a new kind of being" (Christiano 2015: 73).

This response has intuitive appeal. Certain properties like rationality can transform an organism into a new kind of being. Becoming a person transforms a human being's life from merely biological to autobiographical. However, Christiano notes a significant worry: "…it does invoke a difficult metaphysics of essences that may prove intractable" (Christiano 2015: 73). Because persons are *individuals* and not merely individuated, the metaphysical problem is especially challenging.

Consider that the essential properties of some kinds are scalar. The genetic properties constitutive of chimpanzees are not normally scalar, while the properties which constitute persons are scalar. All chimps are equally chimpanzees, but being a person is a matter of degree in two senses: general and particular. The variation in degrees of personhood poses a metaphysical difficulty that does not have a parallel when talking about species membership.

Christiano claims it makes sense to enhance the rational capacities of beings who are already persons and that "…this must be the desire of any rational being.… It is not within the realm of concerns of the

sentient being to become rational" (Christiano 2015: 73). Given the empirical data about connectedness to future selves, many persons do not have such desires and might be incapable of forming them for multiple reasons. Some persons might be cognitively unable to think abstractly enough about what increased rationality is and what it would entail. It appears persons must have a sufficient degree of connectedness to their future selves in terms of similarity, vividness and positivity to have the desire for increased rationality. Thus, it's not in the realm of concern for all rational beings to become more rational. Among those that do have such a desire, some will have it to a much greater degree than others.

Dramatically increasing the rationality of a person might make that person unrecognizable to himself and others. Increased rationality would likely involve an increase in IQ and an increase in the similarity, vividness and positivity a person has toward his future self since these features are necessary for rational self-interest. Imagine the contrary situation where your level of rationality is dramatically reduced to the point you still belong in the class of persons; you're just past the threshold. How different would your life be? Would you still be able to understand and enjoy reading philosophy? Would you be able to have the same meaningful relationships and connections with friends and family? Would you have the same kind of relationship with your current and future selves? We would be right to wonder if you were the same *individual*. We might wonder if a radical change in degree of rationality would destroy a person's essence or eliminate the psychological connections that matter most.

Christiano claims increasing the chimpanzee's rationality beyond a certain point is "not identity-preserving" (Christiano 2015: 73) because the chimp becomes a different kind of being. Increasing a person's level of rationality "… is not to turn the ordinary human into a new type of being" (Christiano 2015: 73). Based on the previous considerations, the claim that increased rationality for persons is identity preserving is likely false assuming the change is dramatic enough. First, increasing a person's rational capacity increases that person's general personhood. The new being is more of a person than the old person. Second, it may be tantamount to a loss of personal identity or a loss of relevant psychological connections. Finally, if the increase in rationality now affords the person type-3 future-directed interests and a much stronger affective connection to future selves, that person can now exercise the capacity for autonomous self-creation. The person has become a new kind of being—a self-creator.

Perring presents a "superperson" who develops features relevant to personhood far beyond the typical human (Perring 1997: 186). Kaufmann comments similarly that Nietzsche maintained "… the gulf separating Plato from the average man is greater than the cleft between the average man and the chimpanzee" (Kaufmann 2013: 151)

The differences between the superperson and the typical person could exceed the differences between a chimp and an average person in terms of degrees of general personhood and particular personhood. The average man may be barely more of a self-creator than a chimpanzee, both being largely products of accidental factors rather than autonomous choices to become a *particular someone*. Treating differences in rationality and self-determination above the threshold as arbitrary differences in talent that should be mitigated by egalitarian considerations—an idea examined by Christiano—is problematic. Rationality and autonomy have an ontological role to play in the kind of beings they constitute even above the threshold; they are essential properties. Mere talents, like other arbitrary factors such as race and gender, have no ontological role to play.

Carter (2011) notes neo-Kantians avoid Kant's non-empirical conception of rationality and instead appeal to a naturalized conception of reason. Carter claims, as if echoing Nietzsche, neo-Kantians and political philosophers fail to address the moral implications of a naturalized conception of reason. Once naturalized, rationality is scalar, and while it may provide a basis for respect, it doesn't necessarily provide a basis for equal respect. The failure to take seriously the problem a naturalized account of reason poses for morality is an example of Nietzsche's claim in *The Antichrist*: despite knowing god is dead "*…everyone nonetheless remains unchanged*" (Nietzsche 1968: 160).

Carter discusses range properties as a basis for equality. He argues there are two issues with the use of range properties. First, we need to know why the range property is morally relevant. If the subvenient property is more fundamental, why shouldn't we focus on that property? Second, if the range property is morally relevant, we would need additional arguments to show the base properties are not relevant. Equally instantiating a range property is not enough to guarantee equality since people would still have the subvenient property in varying degrees.

There is a more fundamental objection to range properties. We need to establish range properties are real properties distinct from their disjunctive base properties. A similar problem occurs in the philosophy of mind with multiply realized higher-order properties.[8] If we ask where a mental property gets its causal powers, the most plausible answer is from its physical base property. Given that a mental property can have multiple physical realizers all with varying causal powers, it is doubtful that mental properties have determinate causal powers of their own that would make them scientifically respectable properties. An empirical property ought to play a causal or metaphysical role independent of our ethical theorizing. Range properties are generally treated as empirical properties, but their ontological status is largely unclarified. We need to show range properties have some causal role to play which is distinct from their base properties. The problem is that they don't have

[8] See Kim 2011: 184–186.

an explanatory role above and beyond what is provided by the base property. All of the causal or metaphysical work that determines degrees of personhood is accomplished by the particular base properties of each individual.

Carter intends to go beyond Rawls' account of range properties by attempting to find "…an independent reason for assessing persons in terms of the range property rather than in terms of the basis of that range property…" (Carter 2011: 550). Of course, if the range property is not a real property, there is nothing to assess persons in terms of other than the base property. When we examine Carter's approach, the only properties ever evaluated are the base properties of particular individuals. Evaluative abstinence does not involve the examination or awareness of a higher-order range property. Carter explains, "…we need to *avoid* looking inside people" (Carter 2011: 551). Avoiding looking at X does not entail looking at Y instead. Choosing not to see the unequal extent to which people have a property does not entail our empirically encountering some other higher-order range property equally instantiated.

Carter admits looking inside a person is a "precondition" for determining whether we should subsequently treat that person as opaque (Carter 2011: 552). We determine if a being is above the "absolute minimum" threshold, but once we have surmised they are above, we adopt an attitude of opacity. Again, we determine whether a being has passed the threshold not by examining a range property but by looking at the degree to which the being has the base property. The range property itself has no role to play in the theory. Furthermore, if Perry's superman or Nietzsche's Plato has on ontological status far superior to the average person which involves differences in kind, choosing to intentionally disregard or ignore their superior status would be disrespectful. It would be disrespectful in a way analogous to failing to treat the average person as being worthy of greater respect than Koko the gorilla.

Carter's argument depends on the appropriateness of opacity respect to beings who meet a certain absolute standard of personhood. He argues that additional considerations about respect override the conclusions we would draw about moral status based on factual differences. Opacity respect is supposed to provide independent, non-question-begging justification for ignoring variations in the base properties. In order for the argument to succeed, it must not implicitly include some notion of equality. Carter gives two necessary conditions for adopting an attitude of opacity respect. The first is when the being "possesses *dignity as agential capacity*," and the second is when it is appropriate for us to view that being *simply as an agent* (Carter 2011: 556).

Carter correctly avoids a mythological, Kantian notion of reason, yet the notion of dignity as agential capacity introduces another mythical concept. Persons vary in their agential capacities and how well they use those capacities. It's unlikely we can make good sense of agency

apart from how it is exercised. The idea that there is a general property of agency that bestows dignity on persons regardless of how it is exercised (i.e., its content) or the degree to which it is instantiated is a way of mythologizing agency. An argument is needed based on an empirical account of human agency that then entails, through the application of a bridge premise, the possession of the moral property of dignity.

Carter focuses on the relationship between political institutions and citizens to justify the second condition. However, how political liberals feel about institutions treating citizens might be another instance of Nietzsche's "everyone nonetheless remains unchanged." Carter explains that political liberals feel the state should not evaluate the rational capacities, abilities to make responsible decisions, or abilities to form worthwhile life plans of its citizens, and that doing so would be disrespectful. In contrast, a Nietzschean might find it disrespectful for the state to treat higher-type persons as equal to those lacking in robust rationality and higher forms of autonomy. Thus, "appropriateness" is a matter of taste and one influenced by our affects and prior moral commitments.

One might object that the commitment to evaluative abstinence is the result of a desire to avoid admitting people have unequal status and therefore they ought to have unequal basic entitlements. Carter responds that the liberal commitment to the outward dignity of agents is not based on the equality of agents but is instead based on "respect for agency itself" (Carter 2011: 558). However, the idea of "agency itself" smuggles in a notion of equality. "Agency itself" is a conceptual abstraction which ignores the real agential differences between persons. As previously mentioned, it is a mythological notion of agency used to justify equality in a way analogous to Kant's non-empirical account of rationality.

Miklosi (2022) argues against the Response Co-variation Thesis: "If there is a valuable property P, such that its presence constitutes a reason for a certain kind of response R to its bearers, then every variation in the degree of P necessarily constitutes a reason for a corresponding variation in R" (Miklosi 2022: 374). The implicit acceptance of this thesis severely limits the argumentative strategies available to advocates of status parity. By rejecting the co-variation thesis, Miklosi hopes to increase the conceptual space for responses to variations in status-grounding properties which would preserve equality. Miklosi argues that although the co-variation thesis may hold for some values, it is not generally true, and importantly it does not hold in the case of valuing rational beings. The same considerations which explain the significance of rationality for moral status in the first place will explain why differences above the threshold don't matter to moral status.

Miklosi examines cases in which the response R does not vary despite variations in property P. These cases are intended to provide an analogical basis to show our responses should not change when

we encounter varying degrees of rationality above the threshold. The strength of this inference depends on two factors: (1) whether Miklosi's description of the initial cases is correct, and (2) whether the latter case involving rationality is sufficiently analogous.

Miklosi uses two initial cases: "historically significant" and "worthy of philosophical understanding" (Miklosi 2022: 378). Once a topic has crossed the threshold of either "historical significance" or "worthy of philosophical understanding," variation in degree of significance or worthiness do not and should not affect the seriousness of our response R. P will vary above the threshold but R will not. Miklosi states the norms which govern intellectual inquiry are not "different or less stringent," and the same norms of "seriousness, devotion, sincerity, clarity, and precision" apply equally (Miklosi 2022: 378). Miklosi admits variations in P might affect whether we will engage with P, but they do not affect *how* we engage with P. Once we determine P is worthy of engagement, changes in P do not affect the norms which govern *how* we engage.

The first objection to Miklosi's account of these cases is that the norms which apply to intellectual inquiry are not, in general, determined by the objects of study. The *basis* of intellectual or academic norms of honesty, precision, rigor, etc., is found in the goals of intellectual activity such as knowledge and, in the case of professional academics, fair evaluation of oneself and one's peers. Therefore, the reason R does not vary with changes in P in these two cases is precisely because R is not based on P to begin with. The intellectual norms which govern our responses to things we find intellectually interesting or significant have little or no grounding relationship to those things. Using Miklosi's distinction, the variations in the objects of study are the grounds for *whether* we engage, but they are not the grounds for *how* we engage.[9]

A second objection challenges Miklosi's intuitions about these cases. The first objection grants we have an equal response R to variations in P, but claims that fact is not relevant because the basis of R is independent of P. The next objection examines whether equal responses are in fact always appropriate. A researcher who lied or misrepresented facts about the holocaust, perhaps by minimizing the atrocities, does something much worse than a researcher who minimizes the extramarital activities of a past government leader. Intuitions might reasonably vary between those who think intellectual norms are equally stringent regardless of subject matter and those who do not.

The first objection lends itself to a third objection which challenges the analogy these cases are intended to provide regarding our response to the value of rational beings. In the case of moral status, the value of rational beings is supposed to *ground* the norms governing our responses to such beings. The moral status of beings is supposed to explain *both* why we should engage with them, and why we should

[9] I believe the same is true of Miklosi's conversational norms.

engage with them equally (i.e., the how question). The earlier cases of historical significance and philosophical understanding are only analogous to moral status regarding the engagement question and not in terms of the norms governing the engagement. Therefore, the fact the co-variation thesis does not hold in those cases suggests very little if anything about moral status of persons because the grounding relation is absent.

Miklosi argues we ought to value rational beings because they are valuers by which he means they are capable of responding to "the reason-giving aspects of the world" (Miklosi 2022: 381). Miklosi continues, "As far as each rational being's own life is concerned, their decisions about which of many rationally eligible goals to adopt should be treated as authoritative. We should treat rational beings as authorities regarding their own lives" (Miklosi 2022: 381). It's not clear what the exact inference is to the conclusion that each and every typical person should be considered an authority about their own lives. Furthermore, it's not clear why we ought to treat people as being *equally* authoritative about their own lives. We typically view persons as authorities when they have extensive or specialized knowledge which then produces the right kinds of results. Given the empirical data mentioned earlier, such as the temporal discounting by those who are less connected to their future selves, some persons are far better than others at adopting goals, appropriately evaluating those goals, and pursuing them rationally. Why then should we treat them as equally authoritative?

Miklosi's justification for why our responses to rational valuers should not "be modulated in a way that tracks variations in levels of rationality" (Miklosi 2022: 382) is interesting. He writes:

> Rational beings are valuers capable of incorporating value in their lives in a distinctive way, i.e., through engaging with it. The crucial point is that it is *only through their own valuing activity* that rational beings can realize this distinctive form of value, which explains the reason for "respect," i.e., treating their own determinations of reason as far as their own life is concerned as authoritative. (Miklosi 2022: 382)

The relationship Miklosi creates between valuing and rationality is important. He claims decisions persons make about "which of many *rationally eligible* [my emphasis] goals" are pursued should be treated as authoritative. Non-rational beings are capable of valuing but that valuing is explained by instinct or emotion. The fact my dog values lying on my couch rather than the floor or a child values a candy bar before dinner ought not to be taken as authoritative. Not all valuing activity is equally valuable, or valuable at all, and therefore does not automatically confer value on the valuer. The respect owed to persons' choices is proportional to the rationality of those choices and the overall degree of rationality of the person.

Even if it is true that we ought to take the rational decisions of persons as decisive, generalizing to the conclusion that we should treat rational beings as authorities regarding their own lives is a *non sequitur*.

Respect for some decisions does not logically entail respect for all decisions which constitute a person's life. Perhaps we ought to respect the short-term choices of persons lacking a significant amount of foresight, but respect less or not at all their long-term choices. Parents of teenagers wisely modulate the degree of autonomy allotted to their children based on the degree of rational decision-making they exhibit.

Somewhat similarly, Carter (2018) argues it is consistent with some egalitarian principles to limit the initial freedom agents have based on their limited degrees of temporal integration. The right to freedom does not have the same "absolute weight" it does when we assume personal identity depends on a non-reductive "further fact" (Carter 2018: 838). Recognizing the empirical fact that persons vary greatly in their temporal unity, he writes: "The kind of agent relevant for a theory of equality of opportunity is not the unified agent but the normal agent. Respect for normal agents is compatible with the enforcement of certain interpersonally uniform limits on the right to distribute future opportunities intrapersonally" (Carter 2018: 838). This kind of egalitarianism is almost Nietzschean. The normal agent has one normative principle of freedom, and the highly integrated agent has another.

Carter admits that were temporal unity constituted by a "further fact" or some other account of "highly diachronically integrated agents" then there would be an "unlimited" right to freedom (Carter 2018: 838). Carter thinks the "further fact" approach is philosophically unsound and the notion of "highly diachronically integrated" agents is "empirically dubious" (Carter 2018: 838). According to Carter, "normal agents" are not that highly integrated. However, given the empirically informed and philosophically motivated account of temporal integration offered here, we do have reasons for thinking some persons have a right to greater freedom than others consistent with some of Carter's own claims.

Returning to Miklosi, there are two additional claims he believes strongly support equal responses to all rational beings: the Directedness Claim and the Singularity Claim. The former claims rational beings themselves are the bearers of value rather than the states of affairs their choices create. The latter claims each rational being has only one life to instantiate value (Miklosi 2022: 382–383). It is important to recall Miklosi previously explained that the same considerations which explain the significance of rationality for moral status in the first place will explain why differences above the threshold don't matter to moral status. In other words, the way in which rationality is relevant to *separating* beings into higher and lower status will also help explain why we should give typical persons equal moral status.

However, Miklosi's summary of the significance of the Directedness and Singularity claims in response to the objection that his view entails we ought to treat minimally rational non-human animals the same way we treat human persons undermines his earlier claims for equality among typical persons. To see why this is so, I quote Miklosi at length:

> In particular, it seems to me that the view suggests an important divide between beings who are responsive to reasons in a way that enables them to make sense of their lives as wholes, to have a broadly temporally extended sense of their own existence that is capable of being organized in response to reasons, on the one hand, and beings who are responsive to reasons in more immediate and localized ways, without a sense that their lives as wholes could hang together on the basis of long-term pursuits and relationships. (Miklosi 2022: 284)

Miklosi relies on the concept of a richly temporally extended sense of self which involves pursuing long-term interests and relationships. The threshold for Miklosi depends on the relative degree of temporal extension or localization.

In terms of my earlier analysis, Miklosi's typical person must have richly integrated type-3 future-directed interests to which the current self has strong affective connections in terms of vividness, similarity and positivity. Unfortunately, the research suggests many people do not have a "broadly temporally extended sense of their own existence" which is "responsive to reasons." Instead, some people view their future selves as strangers and, as we have seen, that difference is reflected in their brain activity. There is no reason to suppose that all persons have robust type-3 future-directed interests or that they are capable of such interests. Thus, the supposed threshold between non-human, minimally rational animals and typical persons carves a further morally significant difference among so-called "typical" persons.

Defenses of moral equality attempt to address the scalar nature of morally relevant properties. These theories correctly note that empirical or naturalized accounts of temporal unity, rationality, and autonomy require us to recognize their scalar nature. However, these theories frequently fail to notice naturalized accounts of those properties also involve recognizing differences in kind which constitute real differences between persons. Attempts to make these properties binary through the use of range properties will not address the variations in kind among persons. While there may be responses to some of the internal challenges to these defenses, different approaches will be needed to address the differences in kind among persons.

References

Arneson, R. 1999. "What, If Anything, Renders All Humans Morally Equal?" In D. Jamieson (ed.). *Peter Singer and His Critics*. Oxford: Blackwell, 103–128.

Carter, I. 2011. "Respect and the basis of equality." *Ethics* 121 (3): 538–571.

Carter I. 2018. "Equal Opportunity, Responsibility, and Personal Identity." *Ethical Theory and Moral Practice* 21: 825–839.

Christiano, T. 2015. "Rationality, Equal Status, and Egalitarianism." In Uwe Steinhoff (ed.). *Do All Persons Have Equal Moral Worth? On "Basic Equality" and Equal Respect and Concern*. Oxford: Oxford University Press, 53–75.

DeGrazia, D. 2007. "Human-Animal Chimeras: Human Dignity, Moral Status, and Species Prejudice." *Metaphilosophy* 38 (2–3): 309–329.

Floris, G. 2019. "On the Basis of Moral Equality: A Rejection of Relation of the Relation-First Approach." *Ethical Theory and Moral Practice* 22: 237–250.

Floris, G. 2020. "Two Concerns About the Rejection of Social Cruelty as the Basis of Moral Equality." *European Journal of Political Thought* 19 (3): 408–416.

Hershfield, H. E. 2011. "Future Self-continuity: How Conceptions of the Future Self Transform Intertemporal Choice." Ann N Y Acad Sci. 2011 October 1235: 30–43.

Husi, S. 2017. "Why We (Almost Certainly) Are Not Moral Equals." *Journal of Ethics* 21: 375–401.

Kaufmann, W. 2013. *Nietzsche: Philosopher, Psychologist, Antichrist*. Princeton: Princeton University Press.

Kierkegaard, S. 2005. "The Rotation Method (from Either/Or)." In R. Solomon (ed.). *Existentialism*. Oxford: Oxford University Press.

Leiter, B. 2019. "The Death of God and the Death of Morality." The Monist 102 (3): 386–402.

Lipshitz, N. 2024. "Binary Properties as the Basis of Equality." *American Philosophical Quarterly* 61 (2): 157–163.

Locke, J. 1979. *An Essay Concerning Human Understanding*. P. H. Nidditch (ed.). Oxford: Oxford University Press.

McMahan, J. 2002. *The Ethics of Killing: Problems at the Margins of Life*. Oxford: Oxford University Press.

Miklosi, Z. 2022. "The Problem of Equal Moral Status." *Politics, Philosophy and Economics* 21 (4): 372–392.

Nietzsche, F. 1961. *Thus Spoke Zarathustra*. Translated by R. J. Hollingdale. New York: Penguin Books.

Nietzsche, F. 1966. *Beyond Good and Evil*. Translated by W. Kaufmann. New York: Random House.

Nietzsche, F. 1967. *Genealogy of Morals*. Translated by W. Kaufmann and R. J. Hollingdale. New York: Random House.

Nietzsche, F. 1968. *The Anti-Christ*. Translated by R. J. Hollingdale. New York: Penguin Books.

Nietzsche, F. 1974. *The Gay Science*. Translated by W. Kaufmann. New York: Random House.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Perring, C. 1997. "Degrees of Personhood." *Journal of Medicine and Philosophy* 22 (2): 173–197.

Rozebloom, G. 2018. "The Anti-inflammatory Basis of Equality." In Mark C. Timmons (ed.). *Oxford Studies in Normative Ethics Volume 8*. Oxford: Oxford University Press, 149–169.

Sangiovanni, A. 2017. *Humanity Without Dignity: Moral Equality, Respect and Human Rights*. Cambridge: Harvard University Press.

Singer, P. 1993. *Practical Ethics*. Cambridge: Cambridge University Press.

Unger, P. 1990. *Identity, Consciousness and Value*. Oxford: Oxford University Press.

van Gelder, J.-L., Hershfield, H. E., and Nordgren, L. F. 2013. "Vividness of the Future Self Predicts Delinquency." *Psychological Science* 24 (6): 974–980.

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.7
Received: November 18, 2024
Accepted: September 15, 2025

Theoretical Sources of Rawls's Justice as Fairness: Kant, Hegel and Mill

JINGHUA CHEN
Guangdong University of Petrochemical Technology, Maoming City, PR China

Rawls regards the liberalism of Kant, Hegel and Mill as important exemplars in the history of the moral and political philosophy of liberalism of freedom. This paper seeks to demonstrate how Rawls draws on these three predecessors. Rawls's theory of justice as fairness has three major components: the original position, the primacy of the basic structure of society, and two principles of justice. I argue that these three elements in Rawls's theory have parallels in the theories of Kant, Hegel and Mill. Firstly, there are essential similarities between Rawls's original position and Kant's Categorical Imperative procedure: justificatory individualism, the Reasonable presupposes and subordinates the Rational, and the combination of moral and realistic considerations. Secondly, Hegel attributes primacy to the state due to the insufficiency of abstract right and morality compared to ethical life and the incompleteness of family and civil society compared to the political state. The special role of the basic structure of society in Rawls's theory of justice draws from Hegel's emphasis on political institutions in realizing freedom. Finally, Rawls's two principles of justice as fairness have roughly the same substantive content as Mill's principles of justice and liberty of the modern world.

Keywords: Justice as fairness; Rawls; Kant; Hegel; Mill.

1. Introduction

Thomas Pogge and Samuel Freeman, in their renowned monographs on Rawls, enlist some significant historical influences on Rawls. Pogge summarizes, Rawls draws inspirations from Aristotle (the Aristotelian principle), Locke (liberal tolerance), Hume (the circumstances of justice), Rousseau (democratic participation and moral education), Bentham and Marx (the focus on social institutions), Mill (arguments for

freedom of thought and conscience), and Sidgwick (reflective equilibrium) (Pogge 2007: 189). Freeman presents a only slightly different list, including Hobbes, Locke, Rousseau, Kant, Hume, Sidgwick, Mill, Hegel, and Marx (Freeman 2007: 14–28).

Among these influences, Kant's relationship with Rawls has stimulated the most significant scholarly interests. The Kantian image of Rawls has been firmly established. Nevertheless, the close link between Hegel and Rawls has attracted increasing attention (Schwarzenbach 1991; Lange 2009; Bercuson 2016; Gledhill 2020). For some scholars, Rawls's theory is "explicitly Kantian, but implicitly Hegelian" (Galston 1982: 512). Given the wave of communitarian criticisms of Rawls in the 1980s, largely deemed Hegelian, this is a notable academic spectacle. More recently, interpreters have examined influences that are not so explicit in Rawls's statements, such as Wittgenstein's influence on Rawls (Reidy 2010; Galisanka 2019; Bok 2017), American Progressivism's influence (Reidy 2022), pragmatism's influence (Botti 2019), Protestant influence (Reidy 2010; Bok 2017; Nelson 2019).

My paper has a different focus, which is on the internal relationships within the family of liberalism of freedom. Rawls aligns his theory of justice with the tradition of liberalism of freedom, the primary exemplars of which are, according to Rawls, Kant, Hegel and J. S. Mill. Rawls remarks that his theory of justice as fairness learns a lot from these three predecessors. The tightly circumscribed primary objective of the paper is to elaborate this remark and examine in detail the internal relationship within the school of liberalism of freedom and put Kantian, Hegelian and Millian elements into the appropriate places within Rawls's theoretical building.

Similar work has not been done yet in the present secondary literature. Since Rawls's Kantianism has been discussed extensively, this paper chooses to spill more ink on Hegelian and Millian aspects. Yet the trio of crucial similarities between Kant's CI-procedure and Rawls's original position has not been shown as a whole by existing literature, including those of Pogge, Freeman, and O'neill. For instance, Pogge argues that Rawls tries to accommodate three criticisms of Kant's view, namely its "practical solipsism," "rigorism," and "austerity" (Pogge 1981: 58–65). However, Pogge contends that the situation of the parties in the original position and the situation of the agent in the CI-procedure cannot be parallel, despite Rawls's insistence on the contrary (Pogge 1981: 49). In this respect, I part with Pogge and side with Rawls by uncovering three essential similarities between these two analytical devices. O'neill focuses on replacing Rawls's idealization approach with her abstraction method in constructing moral and political principles. Her point is that Rawls's Kantian constructivism is not Kantian enough (O'Neill 1998: 210–218). In contrast, my study underscores the essential Kantian attributes of Rawls's constructive device.

While Pogge and O'Neill are critical of Rawls, Freeman is mainly a sympathizer and defender. He points out the parallel between Kant's

idea of the Realm of Ends and Rawls's idea of a well-ordered society (Freeman 2007: 22) and the Kantian origin of Rawls's reflective equilibrium (Freeman 2007: 38). Nevertheless, he claims Rawls's initial drafts of *A Theory of Justice* is influenced insignificantly by Kant and only after *Theory* Kant's influence on Rawls rises (Freeman 2007: 22). By contrast, I attempt to highlight the outstanding Kantian elements in Rawls's original position, a key component of *A Theory of Justice*.

Regarding Rawls's Hegelianism, Sibyl Schwarzenbach's 1991 paper is the first influential piece of literature. She, mainly to rebut the "communitarian" attack on Rawls, specifies three typically Hegelian moments in Rawls's theory of domestic justice: reconciliation as the task of political philosophy, the conception of the political person, and the conception of human community and the state with a special role for freedom and human flourishing (Schwarzenbach 1991: 542–555). Her third point overlaps partially with my discussion of the parallel between Hegel's political state and Rawls's basic structure, yet my exploration is much more detailed and nuanced.

Jeffrey Bercuson discusses the Rousseauvian and Hegelian heritage of Rawls's justice as fairness. He highlights the conception of "robust reasonableness" by exposing the Rousseauvian notion of recognition and self-respect and the Hegelian notion of reconciliation implicit in Rawls's justice as fairness (Bercuson 2014: 5). But he identifies Rawls's Hegelianism after Rawls's political turn (Bercuson 2014: 3-4). My paper explores the Hegelian element in Rawls's primacy of the basic structure of society, starting from *A Theory of Justice*.

In comparison, there is far less literature on the relationship between Mill and Rawls than on Kant or Hegel. Among this relatively sparse literature, Gerald Gaus's 1981 paper stands out. He illuminates the convergence of Mill's and Rawls's liberalism. In particular, he shows that both thinkers defend equal liberty on similar grounds: civic and political liberties are necessary to promote the healthy development of human nature and to avoid pathologies (Gaus 1981: 58–65). Unfortunately, the parallel I explain between Rawls's second principle of justice and Mill's corresponding principles has not been explored by Gaus.

Ruth Abbey and Jeff Spinner-Halev contend that Mill shares similar views about individual autonomy with Rawls, and thus Rawls's distinction between his political liberalism and Mill's comprehensive liberalism is not justified (Abbey and Spinner-Halev 2012: 124). My paper has a different focus: to compare the normative, substantive political principles of Rawls and Mill rather than their accounts of individual autonomy, which belongs to the level of justificatory basis.

The position I hold is a middle ground between two opposing tendencies. On the one hand, it is oversimplified to present a single-dimensional interpretation (simply Kantian or Hegelian) of the nature of Rawls's justice as fairness; on the other hand, a long list of historical and contemporary influences may blur the focal points. My study

shows that Rawls's justice of fairness is almost equally influenced by his three forefathers in the tradition of liberalism of freedom. This article's related and secondary aim is to reveal the relatively neglected Millian face of Rawls's variant of liberalism and put Mill in equal status as Kant and Hegel as historical influences on Rawls. A caveat seems necessary: My aim is to elaborate on Rawls's remarks acknowledging the key influence of these predecessors. My focus is on revealing the relevant crucial parallels to show the Kantian, Hegelian and Millian aspects, respectively. However, it is almost impossible to provide exact and complete evidence to prove that every similarity can be counted as a direct historical influence.

Before entering into the formal exploration, it is worthwhile to first clarify the meaning of the liberalism of freedom. The liberalism of freedom is usually contrasted with the liberalism of happiness. The liberalism of freedom specifies principles of political and civic freedoms as its first principles and accords special priority over other principles, such as utilitarian or perfectionist principles, to some basic liberties: liberty of conscience and freedom of thought, liberties of persons and the free choice of vocation, etc. "Moreover, it assures all citizens adequate all-purpose means (primary goods) so that they can make intelligent use of the exercise of their freedoms" (Rawls 2008: 366; Lange 2009: 103).

In contrast, the first principle of the liberalism of happiness (held by utilitarians) is maximizing the greatest happiness of the greatest number. It is only a happy coincidence for utilitarians to affirm basic freedoms for individuals on utilitarian grounds, for this affirmation is contingent. For utilitarians, there is always the possibility that liberal freedoms will not be confirmed after these consequential calculations. When this happens, utilitarianism is not a liberalism at all (Rawls 2008: 366; Lange 2009:104). Thus, we can see that utilitarianism is a precarious foundation for liberalism.

What liberalism of freedom aspires to do is to ground liberalism more solidly. Both Kant and Hegel staunchly object to a utilitarian justification of a system of rights. Utilitarianism is also the major rival theory Rawls challenges and aims to replace with his justice as fairness. "To say it with a single word, the point of a Rawlsian social order is not the happiness of individuals, but their freedom" (Pogge 2007: 192). This opposition to utilitarianism is characteristic of a liberalism of freedom (Rawls 2008: 343; Lange 2009: 104).

The liberalism of freedom also departs from Lockean liberalism, which grounds liberalism in a social contract aimed at securing the private interests of the contracting parties as atomised individuals. This kind of social contract liberalism has faced harsh criticisms by Hegel and other communitarians. Rawls believes the liberalism of freedom is capable of addressing these criticisms. First, the liberalism of freedom has a notion of a common good, including that of protecting basic civic and political freedoms; second, it recognizes the social rootedness of people within the basic structure of society; third, it acknowledges the

intrinsic value of liberal political institutions; fourth, it affirms collective values, like culture, affection, friendship, and love (Mahlmann and Mikhail 2003: 73, 131).

Lastly, it is worth noting that the essential parallels demonstrated at length in this paper are not found in the existing literature with titles such as "liberalism of freedom," including the 2008 paper by Ragıp Ege and Herrade Igersheim and the 2003 review article by Matthias Mahlmann and John Mikhail. After situating my project in the academic background, we can now turn to substantive discussion, beginning with demonstrating the essential similarities between Rawls's original position and Kant's CI-procedure.

2. Rawls's original position and Kant's Categorical Imperative procedure

Kant's Categorical Imperative procedure (abbreviated as CI-procedure) is a device for moral reflection and tests whether the maxims of actions are permitted morally. The CI-procedure contains four steps:

The first step reads: "I am to do X in circumstances C in order to bring about Y unless Z. (Here X is an action and Y is an end, a state of affairs.)" (Rawls 2007a: 168). The second step is: "Everyone is to do X in circumstances C in order to bring about Y unless Z" (Rawls 2007a: 168). The third step: "Everyone always does X in circumstances C in order to bring about Y, as if by a law of nature (as if such a law was implanted in us by natural instinct)" (Rawls 2007a: 168). The fourth step: "We are to adjoin the as-if law of nature at step (3) to the existing laws of nature (as these are understood by us) and then think through as best we can what the order of nature would be once the effects of the newly adjoined law of nature have had sufficient time to work themselves out" (Rawls 2007a: 168).

2.1 Justificatory individualism

The key characteristic of Rawls's methodology is that the principles of justice for the basic structure of society must be justified to the parties who represent the individuals as free and equal rational persons. Rawls asks us to "keep in mind that the parties in the original position are theoretically defined individuals" (Rawls 1999a: 127; Levin and Levin 1979: 82–87). Through the device of the original position, the principles of justice must be justified to individual persons who are distinct and separate without being conflated into one whole (in contrast with utilitarianism).

Correspondingly, a similar justificatory individualism is also reflected in the CI-procedure. This procedure specifies the content of the moral law from the perspective of reasonable and rational persons as finite beings with moral sensibility and desires (Rawls 2007a: 164). Three elements are linked here: the conception of the person, the rea-

sonable procedure of construction and the moral principles. Kant's CI-procedure aims to justify certain principles based on a particular conception of the person.

As Rawls points out, "The description of the original position resembles the point of view of noumenal selves, of what it means to be a free and equal rational being" (Rawls: 1999a, 225). In particular, the nature of free and equal rational beings is embodied in the argumentative conditions, the combination of which is called the original position. Rawls's original position can be regarded as a procedural interpretation of Kant's conception of autonomy (Rawls 1999a, 226) and justice as fairness has a Kantian root of moral personhood: self-respect and equal respect owed to all (Beatty 1983: 487).

2.2 The Reasonable presupposes and subordinates the Rational

Secondly, the strict priority of the Reasonable over the Rational, or the priority of the right over the good, is a salient trait in both procedures. The Rational means the persons represented in the original position have three regulative interests, including two highest-order interests to achieve and apply two moral powers, that is, moral capacities in the sense of justice and the conception of the good, and one higher-order interest in protecting and advancing their conception of the good as best they can. Rational autonomy means that "the parties are simply trying to guarantee and to advance the requisite conditions for exercising the powers that characterize them as moral persons" (Rawls 2001: 527).

On the other hand, the Reasonable is expressed by moral constraints regulating the rational deliberations of the parties. These constraints include familiar formal conditions on first principles: generality, universality, ordering, finality, publicity, the veil of ignorance, the symmetry of the parties' situation with respect to one another, and the stipulation that the basic structure is the first subject of justice (Rawls 1999a: 126; 2001: 529-530).

The relationship between the Rational and the Reasonable is: "the Reasonable presupposes and subordinates the Rational" (Rawls 2001: 530). The combination and structure of rationality and reasonableness in the original position parallel Kant's proposition of the unity of practical reason: pure practical reason must frame empirical, practical reason in the CI-procedure. "For Kant, merely following your desires represents 'heteronomy,' especially when they conflict with morality. Morality must have priority over my inclinations" (Johnson and Cureton 2022). Put another way, "natural inclinations generally require rational constraint" (Wood 1999: XIV). As Rawls claims, the strict priority of the Reasonable over the Rational, or the priority of the right over the good, is an outstanding feature of Kantian constructivism.

2.3 Combination of Moral and Realistic Considerations

The combination of moral and realistic considerations is the third crucial similarity between Rawls's original position and Kant's CI-procedure. By realistic considerations, I mean evaluating the feasibility and efficiency of the candidate principles by considering empirical theories, historical experience and other relevant general facts.

Rawls's veil of ignorance filters morally irrelevant considerations or prejudices to prevent parties from bargaining with special information to protect their special interests. Nevertheless, the parties behind the veil know general facts and theories, such as psychology, economics, and political science. Including these information is intended to ensure that the chosen principles of justice are realistic in the real social world. And the veil of ignorance draws on Kant's limit on information in the fourth step of the CI-procedure, in which the ideal agents adopt a general perspective with limits of knowledge (Rawls 2007: 175–176).

If a liberal conception of justice is to be realistic, "it must rely on the actual laws of nature and achieve the kind of stability those laws allow, that is, stability for the right reasons" (Rawls 1999c: 12–13). Rawls underscores the notion of strains of commitment, which means that the original parties would not enter into agreements that are impossible or difficult to keep. The selected conception of justice can generate its own support and stability for a well-ordered society based on psychological allegiance. Rawls stipulates that these kinds of general information are available for the deliberative parties to assure the practicality of his justice as fairness.

Sometimes, Rawls is considered more realistic than Kant. For instance, Andrews Reath argues that Kant's Categorical Imperative does not consider empirical data about human beings and the world, which Rawls's original position does (Reath 2015: 218). This remark is understandable since Kant wants to formulate "a metaphysics of morals, which must be carefully cleansed of everything empirical" (Kant 2002: 5). Nevertheless, a deeper review of Kant may suggest a different comprehension from Reath's. "Kant's position is grounded on a distinctive theory of human nature and history, whose importance for Kant's ethics has seldom been appreciated… the neglect of Kant's empirical theory of human nature and history is responsible for most of the misunderstandings of Kant's ethical thought that prevail among its supporters as well as its critics" (Wood 1999: XIII). Rawls's interpretation of the CI-procedure is consistent with Wood's**.** There is a paralleling realistic dimension in Kant's CI-procedure, which takes into account the normal conditions of human life (Rawls 2007a: 167). This feature is reflected by "think through as best we can" in the fourth step of the CI-procedure. In arriving at particular duties of justice and duties of virtue through the CI-procedure, "we rely on certain laws of nature and use various kinds of empirical knowledge about our social world" (Rawls 2007a: 250). Both procedures hold that the choice of the most

reasonable principles is based on the knowledge of all the relevant and true theories on human nature and society, taking into consideration practical limitations and social requirements (Rawls 1999b: 541–543).

After elaborating the crucial similarities between Rawls's original position and Kant's CI-procedure, it seems necessary to inquire into Rawls's transition from Kantian constructivism to political constructivism to clarify the nature of Rawls's appropriation of Kant's moral theory.

2.4 Transition from Kantian constructivism to political constructivism

The problem of Kantian constructivism is that there exist competing and even conflicting ideals of the person, Kantian autonomy cannot be reasonably accepted by many other citizens with different moral or religious views of life. Such a foundation cannot offer a conclusive result on moral or political principles (Brink 1987: 83). This difficulty motivates Rawls to seek a new justificatory ground. Since his milestone 1985 paper "Justice as Fairness: Political Not Metaphysical," Rawls has taken a "political turn" and gradually developed an ingenious theory named "political liberalism." The previous Kantian constructivism has been transformed into political constructivism. Given the fact of reasonable pluralism in liberal democracy, Kantian autonomy is now considered inappropriate for addressing the stability problem. A new basis is found in the "public and shared ideas" (Rawls 1993: 90) deeply rooted in the public political culture of constitutional democracy: the idea of society as a fair system of cooperation and the idea of the person as free and equal, rational and reasonable citizens. Due to the change in justificatory basis, justice as fairness is now a "freestanding view." Rawls believes political constructivism as a method to establish the principles of justice is more capable than Kantian constructivism of striving for broader support from and forging an overlapping consensus among adherents of different comprehensive (religious, moral or philosophical) doctrines. In this way, "the area of agreement throughout society will be sufficiently broad to contribute to stability" (Klosko 1997: 636).

In this paper, the original position I compare with the CI-procedure is a variant of Kantian constructivism. Whether my views can apply to the original position in the fashion of political constructivism is another problem. I believe the three essential similarities demonstrated above are still basically plausible, though the related arguments have to be modified. I cannot expand on this judgment due to limited space. It suffices to offer some preliminary remarks here.

Though, in *Political Liberalism*, Rawls shifts away from the talk of Kantian constructivism to the talk of political constructivism and stops identifying his liberalism as Kantian to be more attuned to the fact of reasonable pluralism, Rawls appropriates key ideas substantially from

Kant's moral constructivism as he works out his version of political liberalism. In a word, "Kantian constructivism makes political constructivism possible" (Tampio 2007: 87–92). The leading idea of political constructivism, like Kantian constructivism, is still to draw a connection between a conception of the person, a procedure of construction, and the principles of justice, or put it more succinctly, to formulate the principles of justice by laying out a procedure of construction. More importantly, though Rawls stresses that the fundamental idea of the person for political constructivism is now that of "the citizen," drawn on public political culture rather than on a comprehensive doctrine, the nature remains the same: free and equal, rational and reasonable. Thus, it can be seen the later political autonomy is difficult to be separated from Kantian autonomy (Tampio 2007: 92). Actually, as early as in "Kantian Constructivism in Moral Theory," Rawls suggests the Kantian conception of the person is implicit or latent in the public political culture in a constitutional democracy (Brink 1987: 78). What Rawls's political turn really does is only to relocate the Kantian origin to a "political" domain. This is why I believe the three crucial similarities revealed above are mainly tenable even after Rawls's well-known political turn.

Because Kant's theory is not sufficiently "political," Rawls turns to Hegel for inspiration to build his theoretical palace. His first use of Hegel is well before the "political turn" at the time of *Political Liberalism*. His first political turn is his primacy of the basic structure of society in *A Theory of Justice*, which draws on Hegel's emphasis on the political state.

3. Rawls's special role for the basic structure and Hegel's emphasis on the state

3.1 The idea of basic structure in Rawls's theory of domestic justice

Regarding Rawls's focus on the basic structure of society, Arash Abizadeh writes, "Indeed, Rawls's thesis that the basic structure is the primary subject of justice is routinely cited as one of his most fundamental and enduring contributions to political philosophy" (Abizadeh 2007: 322). In Rawls's three major works concerning domestic justice, i.e., *A Theory of Justice*, *Political Liberalism* and *Justice as Fairness: A Restatement*, Rawls insists that the basic structure of society is the primary subject of justice. His emphasis on the basic structure of society is a prominent feature of his theory of justice.

What exactly is the basic structure of society? The basic structure is the way in which the major social institutions, including the political constitution and the principal economic and social arrangements, fit together into one scheme and how they distribute fundamental rights and duties and determine the division of advantages that arise through social cooperation (Rawls 1996: 258; 1999a: 6; 2001: 10). It would become easier for us to understand the idea of the basic structure if we

contrast it with particular interactions of individual persons or collectives. Rawls proclaims, "The basic structure is the background social framework within which the activities of associations and individuals take place. A just basic structure secures what we may call background justice" (Rawls 2001: 10). The focus on the basic structure aims to realize background justice directly and only indirectly justice concerning individual social interactions.

It is essential that the basic structure of society consists of fundamental public laws. Hence, it is a legal concept and does not refer to the so-called rules or morality in the non-juridical domains. Moreover, the basic structure is constituted by fundamental rather than subsidiary laws. According to Rawls, the basic structure of society is first "a system of common public law which defines and regulates political authority and applies to everyone as citizens" (Rawls 1996: 265). For comparison, Rawls points out that libertarian doctrine, which is formulated prominently by Robert Nozick, has no special role for the basic structure because it deems the state as one of the private associations rather than a system of common public law.

3.2 Special role of the basic structure of society

The next significant question is why the basic structure is the primary subject of justice or why it has a special role. Rawls puts the question like this: "The problem here is to show why the basic structure has a special role and why it is reasonable to seek special principles to regulate it" (Rawls 1996: 265). He continues to explain that the major reason is that the impact of the basic structure is profound and present from the start. He makes this point most precisely in *Justice as Fairness: A Restatement*, "One main feature of justice as fairness is that it takes the basic structure as the primary subject of political justice. It does so in part because the effects of the basic structure on citizens' aims, aspirations, and character, as well as on their opportunities and ability to take advantage of them, are pervasive and present from the beginning of life. Our focus is almost entirely on the basic structure as the subject of political and social justice" (Rawls 2001: 10).

The basic structure has profound and enduring effects, mainly because it determines the social starting places of all the citizens. An unjust basic structure causes deep inequalities among individuals regarding their initial chances in life. To address the problem of deep inequalities, the principles of social justice must first apply to the basic structure consisting of a political constitution and the major economic and social institutions. And "these principles must nevertheless embody an ideal form for the basic structure in the light of which ongoing institutional processes are to be constrained and the accumulated results of individual transactions continually adjusted" (Rawls 1996: 259).

Rawls explains the significance of the basic structure and its associated background justice by alluding to an economic analogy. Rawls contends that even if the starting points are fair to everyone and every separate and independent economic transaction is just, the accumulated results would tend to be unjust without the guarantee of a just market system. Rawls writes, "Unless this structure is appropriately regulated and adjusted, an initially just social process will eventually cease to be just, however free and fair particular transactions may look when viewed by themselves" (Rawls 1996: 266).

Similarly, injustice in our social world does not solely arise from wrongs in individual interactions or evils in human nature, such as deceit and fraud. The real difficulty is that even if all separate transactions are fair, the overall result would tend to be unjust due to "social trends and historical contingencies" (Rawls 1996: 266).

It is worth emphasizing that Rawls's stress on the basic structure does not mean excluding the significance of justice concerning individual social interactions. Instead, he proposes insightfully "an institutional division of labour between the basic structure and the rules applying directly to individuals and associations and to be followed by them in particular transactions" (Rawls 1996: 269). Rawls argues that with the realization of this division of labour, individual acting agents can be free to pursue their goals more effectively because the functioning of the basic structure regulated by proper principles of justice is always making corrections to offset biased results in the individual transactions (Rawls 1996: 269).

3.3 Hegel's institutional idea of ethical life and his emphasis on the state

Rawls comments that Hegel's institutional idea of ethical life (Sittlichkeit) and his view of persons as rooted in and fashioned by the system of political and social institutions they live in are important contributions to moral and political philosophy. He admits, "*A Theory of Justice* follows Hegel in this respect when it takes the basic structure of society as the first subject of justice" (Rawls 2007a: 336). In the following discussions, I will demonstrate in detail Hegel's institutional perspective concerning human freedom and the relationship between this idea and Rawls's emphasis on the special role of the basic structure of society.

Hegel's most significant work in political philosophy is *Elements of the Philosophy of Right*, which elaborates on the tripartite abstract right-morality-ethical life structure. In the book's first two parts, Hegel discusses abstract rights and morality, both of which embody significant dimensions of human freedom. But they are too abstract and not concrete enough, and hence, unable to stand on their own as independent realities. These two abstract standpoints must be united by ethical life to achieve the concreteness of freedom. Ethical life exists

in three various forms—the family, civil society, and the state (Hegel 1991: 197–198). "In ethical life, we no longer have to do with pure abstractions but with concrete forms of social life" (Franco 1999: 234).

Hegel presents a system of institutions that he believes is the ideal typical social and political system for his country, Prussia, to actualize modern freedom at his age. Hegel opposes absolutism and majority democracy and tries to find a third way. He describes a constitutional monarchy as the most appropriate political system to express and reflect subjective freedom required by the modern world. His constitutional structure comprises three branches, the legislative, executive, and royal powers, which are differentiated but simultaneously united under the monarch, who is checked by the constitution (Franco 1999: 308).

Although Hegel's defence of monarchy, his faith in bureaucracy, and his suspicion of democracy are contrary to the later development of liberal democracy, the interpretation that Hegel is a moderate liberal and defender of the modern constitutional state has become dominant in Hegel scholarship in the past more than a half-century. Knowles sees Hegel as "the greatest and most sophisticated of philosophers of freedom" (Knowles 2002: 27). Rawls interprets Hegel as a liberal thinker and regards his liberalism as an important exemplar in the history of the moral and political philosophy of the liberalism of freedom (Rawls 2007a: 330).

Hegel's construction of political philosophy can be understood as starting from his discontent with classical individualistic liberalism. He believes that earlier liberals, such as Locke and Kant, neglected the profound rootedness of persons within established political and social institutions. They are, to various degrees, "suspicious of the state and saw its structures at best as guarantees for individual liberty, the existence of which was anchored outside the state" (Avineri 1974: 181). In contrast, Hegel asserts that the concept of freedom can only be concretized in the social and political institutions at a particular historical moment.

Ethical life is the key idea Hegel introduced to remedy the individualistic deficiency of classical liberalism. "For Hegel, ethical life is important because it is the realization of our rational essence, freedom" (Franco 1999: 224). Ethical life has absolute authority over the individual; hence, individual freedom is inalienable to a rational ethical life. Freedom of the will is inseparable from social freedom. "It is distinctive of Hegel's thought in these areas that we can act freely only in the context of a form of social life that sustains and protects that freedom; a free society is necessary if freedom of the will is to be a real feature of citizens' lives" (Knowles 2002: 26). According to Hegel, individual freedom can be actualized only within a communal context (Westphal 1993: 234).

Among the three moments of ethical life, the state is the cornerstone. For Hegel, the state, and only the state, is the actuality of con-

crete freedom. And freedom is the rational essence of human beings. Although membership in the family or civil society can confer personal freedoms, these are limited and often in conflict. The state plays a unique and crucial role in actualizing freedom (Knowles 2002: 325).

3.4 Paralleling structures between Hegel and Rawls

It is evident from the discussions above that morality and politics are closely intertwined for both Hegel and Rawls. For Hegel, Kantian autonomy cannot be realized without an ethical community whose institutions are rational and accepted by the citizenry (Gordon 2000: 320). In particular, human freedom, or a fully rational and good life, cannot be fully actualized apart from a rational (reasonable) structure of social institutions. Only within this reasonable and rational social framework can individuals become bearers of culture—religion and philosophy, science and art.

When Rawls stresses that the basic structure of society is the primary subject of justice, he indicates the inseparability of the relationship between constitutional politics and justice. Following Hegel, Rawls insists that background justice of the basic structure of society, determining freedom and equality, is profoundly critical for achieving the citizens' final aims or rational life plans. Both Hegel and Rawls not only respect the significance of civil society but also insist on the primacy of the constitutional system in truly realizing freedom.

Hegel's doctrine of the state is the climax of his political philosophy. The state is the substructure of all the abstract rights, morality and the institutions of the family and civil society. The full development of civil society presupposes the state, and only within the state does the family first develop into civil society (Franco 1999: 25). Just as the categories of right and morality must have the ethical life as their foundation, the complete actualization of freedom needs the transition from family and civil society to the state (Franco 1999: 191). There exists an interdependent relationship between individual liberty and the state. The individual can synthesize the values of family and civil society only as a citizen of the state. Only within the state can all institutions that embody freedom of choice be fully nourished. Human freedom is the telos of the state, and simultaneously, the state is the individuals' aim and purpose (Franco 1999: 278). This is the unification of universality and particularity characteristic of Hegel's philosophy. The stability of society is not simply grounded in the satisfaction of the particular interests of citizens but also in their recognition of the universal interest in maintaining the political and social institutions that make their freedom possible. "Citizens knowingly and willingly acknowledge this universal (collective) interest as their own, and they give it the highest priority. They are ready to act for it as their ultimate end. This is the goal of the project of reconciliation" (Rawls 2007a: 355).

Therefore, we can see that, for Hegel, ethical life, as a synthesis between abstract right and morality, is the true guarantee of subjective freedom. The ethical life consists of non-political spheres, namely family and civil society, and the political sphere, i.e., the state. Hegel advocates pluralism and insists on the necessity of autonomous, voluntary bodies separate from and independent of the state. "While civil society gives existence to the important principle of subjective freedom that distinguishes the modern world from the ancient, by itself it represents only an incomplete actualization of human freedom and one that needs to be distinguished from and subordinated to the full actualization of human freedom in the state" (Franco 1999: 252). Thus can be seen the political state plays a special role in the cause of human freedom. There exists "pronounced primacy of the political" (Avineri 1974: 181) in Hegel's political philosophy.

A similar structure in Rawls's theory of justice is reflected in his institutional division of labour: local, domestic, and global justice. Local justice is imposed directly on human behaviours and dispositions in the non-political sphere, especially within the associations and institutions operating within the framework of the basic structure of society. Domestic justice applies to the basic structure of society, and global justice concerns the principles for international relations and international laws (Rawls 2001: 11). There should be distinct principles of justice for different cases with different aims and natures. Nonetheless, Rawls holds that a proper conception for the domestic basic structure has regulative primacy and is illuminating for the determination of principles of justice in other subjects. Domestic justice indirectly constrains associations and institutions within society, such as churches, universities, companies, clubs, and civil associations. He reminds us that unlike utilitarianism, which is a general and comprehensive conception for all kinds of subjects, justice as fairness is merely a political conception of justice and applies first to the domestic basic structure. In Rawls's theory, the formulation of the principle of global justice and local justice is subordinate to the agreement in domestic justice (Rawls 1996: 262).

To conclude, the special role of the basic structure of society in Rawls's theory of justice draws from Hegel's emphasis on political institutions to realize freedom. After demonstrating the relationship between Rawls and, Kant and Hegel, we continue to show how Rawls is also Millian, to which has not been paid enough attention.

4. Rawls's two principles of justice and Mill's political principles

Rawls writes, "In many of his writings, Mill states certain principles which he sometimes calls 'the principles of the modern world.' These principles we can think of as principles of political and social justice for

the basic structure of society" (Rawls 2007b: 267). These principles aim to protect the rights of individuals and minorities under a democratic regime. Rawls continues, "Now I believe that the content of Mill's principles of political and social justice is very close to the content of the two principles of justice as fairness. This content is, I assume, close enough so that, for our present purposes, we may regard their substantive content as roughly the same" (Rawls 2007b: 267).

It is well-known that Rawls's two principles of justice consist of the first principle, namely the equal basic liberties principle and the second principle, which in turn consists of the principle of fair equality of opportunity and the difference principle. These principles primarily apply to the basic structure of society, govern the assignment of rights and duties, and regulate the distribution of social and economic advantages (Rawls 1999a: 53). In this part, I explore the similarity between Rawls's two principles of justice and Mill's principles of political and social justice and attempt to demonstrate the components in Mill's theory corresponding to the principles in Rawls's justice as fairness, respectively.

4.1 Rawls's First Principle and Mill's Principle of Liberty

The final statement of Rawls's first principle in *A Theory of Justice* reads, "Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all" (Rawls 1999a: 266). Later, in *Justice as Fairness: a Restatement*, Rawls, to respond to H. L. A. Hart's criticism, slightly revises the formulation of the first principle. The first principle now reads, "(a) Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all" (Rawls 2001: 42). The most significant change is the substitute of "a fully adequate scheme of equal basic liberties" for "the most extensive total system of equal basic liberties." These minor alterations do not bother the following comparison between Rawls and Mill's political principles. Nonetheless, both expressions are sufficiently alike.

According to Rawls's first principle, the most extensive or fully adequate scheme of basic liberties should be given to everyone equally. "The only reason for circumscribing basic liberties and making them less extensive is that otherwise they would interfere with one another" (Rawls 1999a: 56). These basic liberties are given by a list of such liberties, such as political liberty, freedom of speech and assembly; liberty of conscience and freedom of thought; freedom of the person, the right to hold personal property and freedom from arbitrary arrest and seizure. "These liberties are to be equal by the first principle" (Rawls 1999a: 53).

All crucial elements in Rawls's first principle are already implicit in Mill's corresponding principle of equal basic rights and liberty, which

can also be called simply the Principle of Liberty, to govern the dealings of society with the individual. This principle stipulates that in the self-regarding sphere of action, the individual is sovereign, absolutely independent of external interference, as long as no harm is caused directly to other members of society (Mill 1977: 223–224). Thus, both Mill's and Rawls's principles are closely related to interpersonal noninterference.

Mill identifies specific domains of human liberty: liberty of conscience, thought and feeling, freedom to express and publish opinions, liberty of tastes and pursuits, and liberty of association among individuals (Mill 1977: 225–226). Rawls admits that justice as fairness follows Mill's Principle of Liberty by listing certain enumerated liberties as legal and moral rights of justice, without defining liberty in general or as such (Rawls 2007b: 288).

For Rawls, freedom is essentially equal freedom. Mill also stresses this point. For Mill, the liberties and rights listed above must be extended to all, "no one being now left out" (Mill 1984: 294). He insists that one of the main foundations of modern life is respect for each other's rights. Mill even declares, "The moral regeneration of mankind will only really commence, when the most fundamental of the social relations is placed under the rule of equal justice, and when human beings learn to cultivate their strongest sympathy with an equal in rights and in cultivation" (Mill 1984: 336). For him, equality, as freedom, is the most important political value. Equal freedom is the primary characteristic of the modern world, starkly distinguished from the relationship of domination and subjection in old age.

The priority of liberty in Rawls's theory of justice is also found in Mill's liberalism. Mill asserts, "No society in which these liberties are not, on the whole, respected, is free, whatever may be its form of government; and none is completely free in which they do not exist absolute and unqualified" (Mill 1977: 226). As Riley interprets, "Mill's liberal utilitarian scheme of equal rights is also distinctive because of the absolute protection afforded to the individual's liberty to choose as he likes with respect to certain 'purely self-regarding actions' said to directly cause no 'perceptible damage' to other persons against their wishes" (Riley 1998: 297). Because of this priority of liberty in Mill's political philosophy, Rawls believes that Mill is an important exemplar in the history of the political philosophy of the liberalism of freedom. However, given that Rawls's theory also belongs to the liberalism of freedom, the priority of political and civic freedoms is common among these four liberal thinkers. The similarity that makes Rawls's two principles of justice closer to Mill rather than Kant and Hegel lies in the resemblance between Rawls's second principle and Mill's paralleling doctrine, which we will discuss in the following section.

4.2 Rawls's second principle and Mill's principles of equality of opportunity and egalitarian economic distribution

Rawls's second principle reads, "Social and economic inequalities are to be arranged so that they are both: (a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and (b) attached to offices and positions open to all under conditions of fair equality of opportunity" (Rawls 1999a: 266). In *Justice as Fairness: a Restatement*, the second principle reads, "Social and economic inequalities are to satisfy two conditions: first, they are to be attached to offices and positions open to all under conditions of fair equality of opportunity; and second, they are to be to the greatest benefit of the least-advantaged members of society (the difference principle)" (Rawls 2001: 42). Except for dropping "consistent with the just savings principle" in the difference principle, the statement of the second principle remains the same.

While the first principle applies to the constitutional essentials of the social structure to secure equal basic liberties, the second principle, in order to specify and establish social and economic inequalities, applies to the distribution of income and wealth and to the design of organizations that make use of differences in authority and responsibility (Rawls 2001: 53). "Now the second principle insists that each person benefit from permissible inequalities in the basic structure" (Rawls 2001: 56).

As mentioned above, Rawls's second principle comprises the principle of fair equality of opportunity and the difference principle. Both these two parts have their similar counterparts in Mill's theoretical construction. As Rawls points out in his lectures on the history of political philosophy, Mill also advocates equality of opportunity. Mill distinguishes modern and pre-modern doctrines. Pre-modern society was constituted on the principle of inequality. "All were born to a fixed social position, and were mostly kept in it by law, or interdicted from any means by which they could emerge from it" (Mill 1984: 273). There existed shocking inequality between white and black, men and women, commoners and noblemen, slaves and freemen. Birth could decide a person's position throughout life. Lower status of birth would interdict people from more elevated social positions and respectable occupations. By contrast, according to Mill, the characteristic of the modern world displayed in modern institutions, modern social ideas, and modern life is that "human beings are no longer born to their place in life, and chained down by an inexorable bond to the place they are born to, but are free to employ their faculties, and such favourable chances as offer, to achieve the lot which may appear to them most desirable" (Mill 1984: 272–273). Both Mill and Rawls believe that a just social system should try to eliminate the morally arbitrary factors in determining the life prospects of members of society. However, the account of the principle of equality of opportunity is the most brief among all the principles. In comparison, they spend more time on the principle of distributive

justice. For Rawls, it is called the difference principle; for Mill, the principle of egalitarian economic distribution.

As Philippe van Parijs comments, "Few components of John Rawls's political philosophy have proven so epoch-making as what he somewhat oddly called the 'difference principle.' None has exercised as great an influence outside the circle of academic philosophers" (Parijs 2003: 200). Although there are different interpretations of this principle, the core of the principle is relatively clear: social and economic inequalities ought to be evaluated and justified in terms of how well they advance the interests of the least advantaged. Parijs claims, "The idea of using the latter as the benchmark for assessing inequalities had never been given, before Rawls, a powerful explicit formulation that could capture the scholarly imagination" (Parijs 2003: 200). Nonetheless, the basic spirit of difference principle echoes Mill's principle of egalitarian economic distribution, which makes Rawls's principles of justice closer to Mill instead of Kant and Hegel.

The content of Mill's principle of egalitarian economic distribution can be summarised as follows, "existing competitive capitalism might eventually be transformed into a more cooperative type of private property economy, involving much less inequality in the distribution of wealth than hitherto observed" (Riley 1998: 294). The well-ordered society regulated by Rawls's principles of justice reminds us of Mill's ideal stationary state, which is in essence "a system of universal equal rights accompanied by substantial economic equality" (Riley 1998: 320). In Mill's liberal egalitarian Utopia, namely, a "more cooperative and egalitarian form of capitalism" (Riley 1998: 320), the social goal of a more equitable distribution of wealth for the given population is more crucial than economic growth. Thus, it can be seen that Mill's distinctive brand of liberal utilitarianism is close in spirit to Rawlsian non-utilitarian liberalisms, which give similar prominence to certain equal rights and fair distribution of wealth (Riley 1998: 326–327).

Mill is concerned about the welfare of the working classes and the huge gap between the poor majority and the wealthy minority. He observes that the industrious classes are poor, whereas the idle people are rich. Their life prospects are largely determined by the status they were born into. Great poverty among the working class has little to do with desert. He thus shares similar concerns and grievances with socialists. Nonetheless, he does not share their solutions to these social evils. He has no time for the revolutionary centralized socialists, who want to transform society radically and take over and manage all the property immediately (Ten 1998: 391). Although Mill sympathizes with gradual, decentralized socialism advocated by Owen and Fourier, which is applied to villages or townships on an experimental basis, he pays more attention to reforming capitalism in an egalitarian direction. Following his predecessor Bentham, Mill's liberal utilitarianism "associates the increase of general welfare (and its chief ingredient,

security) with extension of basic rights and reduction of economic inequality" (Riley 1998: 320). Economic equality must be one of the social goals. For Mill, a government that cannot promote equality, whenever this can be done without undermining reasonable private property, only serves the minority and harms the majority and is essentially a bad government (Riley 1998: 320).

Mill also insists that admission to the franchise and the acquisition of "purely political rights" by the working classes, which he ardently supports, are not enough to eliminate the social injustices. To realize substantive freedom, opportunity and development, the system of economic distribution and property institution must be reformed (Ten 1998: 388–389). Likewise, Rawls's difference principle seeks to counter the objection by radical democrats and socialists that the citizens' basic rights and liberties in a modern democratic state are, in practice, merely formal against the background of enormous social and economic inequalities (Rawls 2001: 148). As Rawls writes, "The difference principle, in maximizing the index available to the least advantaged, maximizes the worth to them of the equal liberties enjoyed by all" (Rawls 2001: 149).

Rawls holds not only that his two principles of justice as fairness have the same substantive content as Mill's principles of justice and liberty in the modern world but also that both of them support similar basic institutions of the well-ordered society (Rawls 2007b: 297). The core of Mill's egalitarian project is to challenge and reform the social organization and the established system of private property according to equitable principles to foster a more egalitarian distribution of wealth. The concrete policies include progressive taxation of estates and ensuring reasonable access to natural resources to minimize inequalities of opportunity on the premise that the producer's rights to the fruits of his labour and savings are guaranteed (Riley 1998: 320–321). "These and other reforms of the existing idea of property would tend to promote a far more egalitarian distribution of wealth without subverting capitalism itself" (Riley 1998: 319–320).

Correspondingly, Rawls believes a property-owning democracy, which is different from laissez-faire capitalism or welfare-state capitalism, is the institutional content that realizes the two principles of justice in its basic system (Rawls 2001: 135–136). A property-owning democracy aims to disperse the ownership of productive assets and human capital to put all citizens on a footing of a suitable degree of social and economic equality and avoid developing a discouraged and depressed underclass who feels left out and does not participate in the public affairs (Rawls 2001: 139-40).

Finally, even Rawls's key argument for the difference principle is similar to Mill's reason for his principle of egalitarian economic distribution. In *Justice as Fairness*, Rawls admits that the maximin rule is invalid in arguing against the restricted utility principle favouring

the difference principle. That is because the minimum in the restricted utility principle, as expressed in a capitalist welfare state, will prevent the least advantaged from experiencing their condition as so miserable that they reject this conception of justice. Nevertheless, Rawls contends that the difference principle is still better than the restricted utility principle because it prevents the least advantaged from withdrawing from the political society and becoming passive citizens. It enables them to see themselves as full members of the political society (Rawls 2001: 129–130). Correspondingly, it is well-known that avoiding being passive citizens is the Archimedean point in Mill's theoretical building of political philosophy.

5. Conclusion

I do not mean Rawls has nothing distinctive from his predecessors. It is not difficult to give examples of Rawls's uniqueness. For instance, the principles of justice chosen in Rawls's original position are to regulate the basic structure of society. "By contrast, Kant's account of the Categorical Imperative applies to the personal maxims of sincere and conscientious individuals in everyday life" (Rawls 2007a: 553). Therefore, Kant and Rawls proceed in opposite directions. Kant starts from the particular case in daily life. He believes that numerous interconnected correct personal maxims would eventually constitute a system of moral principles, including principles of social justice, thereby creating a good society. Rawls begins with principles of social justice that regulate the basic structure of society, thereby providing a background of justice and a legal framework for personal and associational activities (Rawls 2007a: 552–553). This discrepancy suggests the Hegelian dimension of Rawls's justice as fairness or "Rawls's Hegelian reading of Kant" (Gledhill 2020: 128).

When it comes to the comparison with Hegel, Rawls emphasizes the distributive part of the basic structure of society more than Hegel for actualizing freedom and justice. Rawls regards the institution of family as a part of the basic structure of society, while Hegel considers it as beyond the domain of the political state. Rawls's idea of the state significantly differs from Hegel's in some other aspects. Hegel asserts the two traditional powers of sovereignty: the state's right to go to war in the rational pursuit of its national interests and its complete internal control over the population. Hegel regards these rights as essential for the idea of the state, which he conceives as a substantive individual or spiritual substance (Rawls 2007a: 360–361). In contrast, Rawls believes that states have no absolute sovereignty. State sovereignty, internal and external, should be constrained by international laws. In addition, the metaphysical idea of "Geist" is crucial in Hegel's theory of the state. For Hegel, only when social and political institutions realize the good and freedom of individuals does Geist achieve its full expression and conscious self-awareness (Rawls 2007a: 370). Rawls

does not rely on the mystical concept of "Geist" in his primacy of the basic structure of society.

Finally, regarding the relationship with Mill, Rawls also has his own distinctiveness. He explicitly expresses the priority rules in his two principles of justice. The first priority rule is the priority of liberty, and the second priority rule is the priority of justice over efficiency and welfare. (The second principle of justice is lexically prior to the principle of efficiency and to that of maximizing the sum of advantages, and fair opportunity is prior to the difference principle) (Rawls 1999a: 266–267). Although similar priority rules might be inferred from Mill's political theory, he does not give a clear presentation and enough emphasis. In addition, there is little analogous to Rawls's detailed discussions of the basic structure of society in Mill's thought. Mill's various political principles are scattered in different works. And he does not stress his political principles as the content of justice (Abbey and Spinner-Halev 2012: 131). In contrast, Rawls's presentation of justice as fairness is much more unified and well-directed in connecting the basic structure, substantive principles and the conception of justice.

In later works, Rawls distinguishes his political liberalism from comprehensive liberalisms, represented by Kant and Mill (I would add Hegel). Due to limited space, this paper can not cover this significant development or family dispute within the liberalism of freedom. However, it is fair to say that Rawls draws substantially from the preceding important exemplars of liberalism of freedom in the history of political philosophy: Kant, Hegel and Mill. And his originality and greatness in his theory of justice as fairness may lie in his indefatigable improvement on and ingenious synthesis of these great predecessors' liberal theories.

References

Abbey, R., and J. Spinner-Halev. 2013. "Rawls, Mill, and the Puzzle of Political Liberalism." *The Journal of Politics* 75 (1): 124–136.

Abizadeh, A. 2007. "Cooperation, Pervasive Impact, and Coercion: On the Scope (Not Site) of Distributive Justice." *Philosophy & Public Affairs* 35 (4): 318–358.

Allison, H. 1996. *Idealism and Freedom: Kant's Theoretical and Practical Philosophy*. Cambridge: Cambridge University Press.

Avineri, S. 1974. *Hegel's Theory of the Modern State*. Cambridge: Cambridge University Press.

Beatty, J. 1983. "The Rationality of the 'Original Position': A Defense." *Ethics* 93 (3): 484–495.

Bercuson, J. 2014. *John Rawls and the History of Political Thought: The Rousseauvian and Hegelian Heritage of Justice as Fairness*. London: Routledge.

Brink, D. 1987. "Rawlsian Constructivism in Moral Theory." *Canadian Journal of Philosophy* 17 (1): 71–90.

Bok, P. M. 2017a. "'The Latest Invasion from Britain': Young Rawls and His Community of American Ethical Theorists." *Journal of the History of Ideas* 78 (2): 275–285.

Bok, P. M. 2017b. "To the Mountaintop Again: The Early Rawls and Post-Protestant Ethics in Postwar America." *Modern Intellectual History* 14 (1): 153–185.

Botti, D. 2019. *John Rawls and American Pragmatism: Between Engagement and Avoidance*. Lanham: Rowman & Littlefield.

Dworkin, R. 1973. "The Original Position." *The University of Chicago Law Review* 40 (3): 500–533.

Ege, R., and H. Igersheim. 2008. "Rawls with Hegel: The Concept of 'Liberalism of Freedom.'" *The European Journal of the History of Economic Thought* 15 (1): 25–47.

Franco, P. 1999. *Hegel's Philosophy of Freedom*. New Haven and London: Yale University Press.

Freeman, S. 2007. *Rawls*. London: Routledge.

Gališanka, A. 2019. *John Rawls: The Path to a Theory of Justice*. Cambridge, MA: Harvard University Press.

Gaus, G. 1981. "The Convergence of Rights and Utility: The Case of Rawls and Mill." *Ethics* 92 (1): 57–72.

Galston, W. A. 1982. "Moral Personality and Liberal Theory: John Rawls's 'Dewey Lectures.'" *Political Theory* 10 (4): 492–519.

Gledhill, J. 2020. "Rawls's Post-Kantian Constructivism." In *Hegel and Contemporary Practical Philosophy*. London: Routledge, 128–152.

Gordon, R. H. 2000. "Modernity, Freedom, and the State: Hegel's Concept of Patriotism." *The Review of Politics* 62 (2): 187–212.

Guyer, P. 2000. *Kant on Freedom, Law, and Happiness*. Cambridge: Cambridge University Press.

Hegel, G. W. F. 1991. *Elements of the Philosophy of Right*. Edited by Hugh Barr Nisbet. Cambridge: Cambridge University Press.

Johnson, R., and A. Cureton. 2022. "Kant's Moral Philosophy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/fall2022/entries/kant-moral/.

Kant, I. 1999. *Kant: Practical Philosophy*. Edited by Mary J. Gregor. Cambridge: Cambridge University Press.

Kant, I. 2002. *Groundwork for the Metaphysics of Morals*. Edited by J. B. Schneewind. New Haven: Yale University Press.

Kaufman, A. 2012. "Rawls and Kantian Constructivism." *Kantian Review* 17 (2): 227–256.

Klosko, G. 1997. "Political Constructivism in Rawls's Political Liberalism." *American Political Science Review* 91 (3): 635–646.

Knowles, D. 2002. *Hegel and the Philosophy of Right*. London and New York: Routledge.

Korsgaard, C. 1996. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

Levin, M., and M. Levin. 1979. "The Modal Confusion in Rawls' Original Position." *Analysis* 39 (2): 82–87.

Lange, M. M. 2009. *Defending a Liberalism of Freedom: John Rawls's Use of Hegel*. New York: Columbia University Press.

Mahlmann, M., and J. Mikhail. 2003. "The Liberalism of Freedom in the History of Moral Philosophy." *Archives for Philosophy of Law and Social Philosophy* 89: 122–132.

Mill, J. S. 1977. "On Liberty." In *Collected Works of John Stuart Mill*, vol. XVIII: *Essays on Politics and Society Part I*, edited by John M. Robson. London: Routledge.

Mill, J. S. 1984. "The Subjection of Women." In *Collected Works of John Stuart Mill*, vol. XXI: *Essays on Equality, Law, and Education*, edited by John M. Robson. London: Routledge.

Nelson, E. 2019. *The Theology of Liberalism: Political Philosophy and the Justice of God.* Cambridge, MA: Harvard University Press.

O'Neill, O. 1989. *Constructions of Reason: Explorations of Kant's Practical Philosophy.* Cambridge: Cambridge University Press.

Parijs, P. van. 2003. "Difference Principles." In *The Cambridge Companion to Rawls*, edited by Samuel Freeman. Cambridge: Cambridge University Press.

Pogge, T. W. 1981. "The Kantian Interpretation of Justice as Fairness." *Zeitschrift für philosophische Forschung* 35 (1): 47–65.

Pogge, T. 2007. *John Rawls: His Life and Theory of Justice.* Oxford: Oxford University Press.

Rawls, J. 1993. *Political Liberalism*. New York: Columbia University Press.

Rawls, J. 1996. *Political Liberalism*. Expanded edition. New York: Columbia University Press.

Rawls, J. 1999a. *A Theory of Justice.* Revised edition. Cambridge, MA: Harvard University Press.

Rawls, J. 1999b. "Kantian Constructivism in Moral Theory." In *Collected Papers*, edited by Samuel Freeman. Cambridge, MA: Harvard University Press.

Rawls, J. 1999c. *The Law of Peoples: With 'The Idea of Public Reason Revisited'.* Cambridge, MA: Harvard University Press.

Rawls, J. 2001. *Justice as Fairness: A Restatement.* Edited by Erin Kelly. Cambridge, MA: The Belknap Press of Harvard University Press.

Rawls, J. 2005. *Political Liberalism*. Expanded edition. New York: Columbia University Press.

Rawls, J. 2007a. *Lectures on the History of Moral Philosophy.* Edited by Barbara Herman. Cambridge, MA: Harvard University Press.

Rawls, J. 2007b. *Lectures on the History of Political Philosophy.* Edited by Samuel Freeman. Cambridge, MA: The Belknap Press of Harvard University Press.

Rawls, J. 2008. *Lectures on the History of Political Philosophy.* Cambridge, MA: Harvard University Press.

Reath, A. 2015. "The 'Kantian Roots' of the Original Position." In *The Original Position*, edited by Timothy Hinton. Cambridge: Cambridge University Press.

Reidy, D. 2010. "Rawls's Religion and Justice as Fairness." *History of Political Thought* 31 (2): 309–344.

Reidy, D. 2022. "Rawlsian Liberalism and/as American Progressivism." *Biblioteca della Libertà* 57: 223–246.

Riley, J. 1998. "Mill's Political Economy: Ricardian Science and Liberal Utilitarian Art." In *The Cambridge Companion to Mill*, edited by John Skorupski. Cambridge: Cambridge University Press.

Rostbøll, C. 2011. "Kantian Autonomy and Political Liberalism." *Social Theory and Practice* 37 (3): 341–364.

Scheffler, S. 1979. "Moral Independence and the Original Position." *Philosophical Studies* 35 (4): 397–403.

Schwarzenbach, S. 1991. "Rawls, Hegel, and Communitarianism." *Political Theory* 19 (4): 539–571.

Tampio, N. 2007. "Rawls and the Kantian Ethos." *Polity* 39 (1): 79–102.

Ten, C. L. 1998. "Democracy, Socialism, and the Working Classes." In *The Cambridge Companion to Mill*, edited by John Skorupski. Cambridge: Cambridge University Press.

Westphal, K. 1993. "The Basic Context and Structure of Hegel's Philosophy of Right." In *The Cambridge Companion to Hegel*, edited by Frederick C. Beiser. Cambridge: Cambridge University Press.

Wood, A. 1999. *Kant's Ethical Thought*. Cambridge: Cambridge University Press.

Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026

Book Review

Alex Madva, Daniel Kelly, and Michael Brownstein, Somebody Should Do Something: How Anyone Can Help Create Social Change*, Cambridge: MIT Press, 2025, 352 pp.*

Picking up recycling or riding a bike to work can often make us feel better about our life choices. It can also, however, make us question how much those choices matter in a world that seems to constantly be in crisis. Sorting my trash only to see it all end up in the same landfill, or giving up driving in a town where I can't tell fog from smog, can make my efforts feel insignificant, especially if I care about climate change and want to avoid being part of the problem. It then becomes easier to consider how such large problems should be tackled by systems and institutions capable of making truly impactful changes. At the beginning of their timely and aptly titled book, *Somebody Should Do Something: How Anyone Can Help Create Social Change*, the authors Michael Brownstein, Alex Madva, and Daniel Kelly directly address the "either/or" mindset that hinders meaningful social change. They reject the false dichotomy between individual and systemic responsibility. Instead of thinking that either individual choices or social-structure changes are the key to fighting structural injustice, they propose a "both/and" approach based on the idea that personal and structural changes are deeply connected. In other words, we should both make individual decisions that drive the change of current systems, while also creating systems that enable people to make better individual decisions.

To demonstrate that resisting injustice should be viewed from a "both/and" perspective, the authors focus on two of the so-called "everything problems," specifically systemic racism and climate change. These problems can't be solved by "fixing" just one issue because many interconnected causes support their persistence. Burning fossil fuels, deforestation, overconsumption, and food production are all factors that worsen the negative effects of climate change. Similarly, the recurrence of numerous obstacles created within a system, including segregation, employment opportunities, education, access to healthcare, and housing, sustains racial discrimination and intolerance. Therefore,

solving a problem that involves various social structures, such as laws, political economies, institutions, norms, and cultures, requires a structural change.

Brownstein, Madva, and Kelly, who are philosophy professors engaged with topics at the intersection of ethics, moral theory, social change, and cognitive science, discuss the concept of structural change as encompassing both large, far-reaching transformations and a process of small, incremental changes that can eventually add up. This perspective further challenges the "either/or" model of thinking by arguing that fighting injustice depends on both individual resistance and institutional reform. So, for the authors, the change doesn't come from "within us" or "the outside of us," but rather results from ongoing feedback between people and systems. Throughout the book, they use clear writing to communicate their ideas, present arguments grounded in social psychology and sociology, and share an inspiring message: anyone can become a changemaker. In the following sections, I will present the case the authors make for their "both/and" approach, exploring the theoretical framework that offers a hopeful and especially optimistic view of initiating and sustaining social change.

Most of us probably know the saying that comparison is the thief of joy. However, it seems that comparing ourselves or our situation to others is a universal part of the human experience. Keeping that in mind, the first step in adopting a "both/and" mindset that the authors suggest is to stop comparing our individual actions to what, for example, a city government can achieve. Feeling like our actions don't matter enough is inevitable when we compare suggesting to our friends that they try riding a bike to work with a city's financial incentive for everyone who uses city bikes as their main mode of transportation. To avoid this feeling, we should compare the efforts of different scales separately. We should also, as Brownstein et al. explain, think in terms of "bundles," that is, a collection of individual choices that can trigger structural changes which, in turn, can enable impactful personal agency. Again, this should foster productive people-system feedback based on understanding that the impact of an individual action depends on the structure that shapes it, while the possibility of structural change depends on the individuals who create and maintain it.

This relationship between the environment and the individuals embedded within it forms a pathway toward the authors' idea of "unlearning habits," which can be applied to social biases at the core of race-based intolerance. As creatures of habit, influenced by the culture, system, and people around us, we often unconsciously conform to prevailing beliefs and behaviors. Just as learning them does, unlearning harmful habits involves shaping new social norms through interactions in which people's actions "signal" different behaviors to others, who then interpret these signals and modify their own behavior. Seemingly small individual acts in a dynamic social context can, therefore, serve

as signals that influence others' choices. As Brownstein et al. further argue, these small acts can accumulate over time and become cascades that reach a "tipping point," a threshold that, once crossed, triggers rapid and significant structural change. However, it is difficult to predict when and how such a change will happen. We might be approaching a tipping point right now, but we might as well never cross one again. Depending on one's perspective, this uncertainty can either be quite discouraging or motivate us to join a promising social movement that could perhaps trigger the next butterfly effect.

In addition to recognizing that personal agency is crucial for signaling the need to challenge existing norms, the authors emphasize the importance of developing "structure-facing skills," which underpin specific social practices aimed at bridging the gap between individuals and the system. These practices stem from an understanding of how social structures work and from acknowledging the concrete steps we could take to tackle a specific issue. Some of the many structure-facing skills that people can learn are, for instance, intersectional awareness, counterfactual thinking, cognitive flexibility, and abandoning political hobbyism. People with these skills tend to be less judgmental and prejudiced, making them more willing to work with others and build alliances with those who have different viewpoints but share common struggles and goals. They are also more likely to understand that "everything problems" require a multifaceted approach, to imagine different scenarios and outcomes when thinking about solving problems, and to feel more motivated to become politically active and involved, rather than treating politics as a hobby discussed with strangers online from the comfort of their own homes. Dedicating one's time and effort to developing structure-facing skills is, therefore, another both/and endeavor that offers a perspective enabling people not only to see how their choices relate to structural mechanics but also to understand it from multiple viewpoints. This can further reveal different roles one might assume on the way to meaningful change.

Recognizing an opportunity to act from where you are and practicing structure-facing skills, whether by joining a labor union, participating in organized boycotts, or publicly advocating for policy change, also involves acknowledging that the results of our efforts won't happen overnight, that their success is always possible but never guaranteed, and that persistence, though it may be tedious, is key to igniting, implementing, and sustaining change. Once all of this comes together in a both/and mindset, the philosophical significance of Brownstein, Madva, and Kelly's fundamental arguments becomes clear: no one is expected to single-handedly "fix" the entire system, but rather to help create the conditions for overcoming systemic inertia.

There is no denying that this theoretical framework is compelling; it offers an invaluable guide away from the "either/or" perspective, emphasizing the potential of each personal choice to develop into a

cascade. Still, even though the authors are aware of the material and psychological constraints that keep people passive, I believe their optimism about the importance of individual acts rests on a somewhat idealized view of human agency that overlooks how difficult it is to become willing to *do something*. Although our capacity to initiate change may seem limited compared to institutions, shifting our focus to individual actions can often lead us to compare our resources with those of our neighbors. The issue then shifts from an either/or mindset to thinking, "I could, but someone else might do it better." If a person next door has a higher-paying job, no children, and more leisure time overall, a tired working parent who has given in to political hobbyism might believe they are "doing their part," expecting their neighbor to do more, since the neighbor appears to be in a better position to do so.

Moreover, even if someone in a better position decides to take that first step and signals to others that it is time to act, motivating them to, for example, join town council meetings, only to see that nothing they propose is ever realized, staying persistent might begin to seem like a luxury many lack the resources to continue investing in. Not feeling discouraged by a lack of results and working to develop structure-facing skills that would prevent us from this kind of linear thinking is challenging, especially when many people operate this way. We are taught from an early age that hard work leads to rewards, whether it's a good grade, more job opportunities, a raise, or the ability to afford a home. Breaking out of that cycle of thinking requires cognitive energy often spent on tasks and responsibilities that secure our livelihood. Before changing the system, a person should first carefully consider what they could modify in their life to become a potential changemaker, a task that is often quite demanding in itself.

Facing structural injustice and feeling motivated to find your role so you can participate in a feedback loop between your community and the system becomes even more difficult if your surroundings are not responsive to your plea. For many people, the need to belong outweighs the urge to speak up for change. When someone's only support network, whether it's family, friends, or coworkers, doesn't see the need to question the status quo, expressing disagreement can mean risking social isolation. Considering that possibility, along with the aforementioned contextual obstacles, adds another reason why individuals might feel stuck in their efforts. Furthermore, the authors' optimism about cascades and tipping points may overlook how resilient entrenched social structures are to attempts to disrupt the current system. While these structures are not immune to meaningful signals, they often benefit more from maintaining the status quo, accepting small compromises that enhance people's sense of agency without destabilizing the power dynamics. By offering inexpensive, limited solutions and few concessions, the system resets the feedback loop, silences the movement, and leaves activists exhausted but with a fleeting sense of achievement.

Under these conditions, the both/and approach proposed by Brownstein, Madva, and Kelly reaches a point of vulnerability: the central question becomes not only how to ignite collective action, but also how to protect that fragile flame from being extinguished by a wary, risk-averse community or a system that only pretends to evolve.

In the end, *Somebody Should Do Something* provides its readers with an indispensable psychological and philosophical guide for dealing with the overwhelming complexity of "everything problems." By challenging the false dichotomy of the either/or mindset, Brownstein, Madva, and Kelly effectively argue that individual actions are not trivial, but essential signals within a dynamic, interdependent system. Their framework issues a loud call to action to those seeking to bridge the gap between personal choices and public transformation. However, as this review has suggested, while serving as a powerful blueprint for how change could occur, the authors' both/and approach assumes a certain level of stability and resources that systemic injustice often undermines. Ultimately, what makes the book so compelling also reveals its main tension: it shows that *anyone* can help create social change, but implicitly suggests that not *everyone* can afford to take the first risk. By identifying the tools for structural change, the authors equip those positioned to act. What remains is to distribute responsibility more evenly so that a both/and mindset becomes standard practice rather than a privilege of the resilient.[1]

MONIKA ZEBA
Institute of Philosophy, Zagreb

[1] This work has been supported by the Croatian Science Foundation under the project No. IP-2022-10-5341.