UNIVERSITA' VITA SALUTE SAN RAFFAELE

CORSO DI DOTTORATO DI RICERCA
IN FILOSOFIA

CURRICULUM IN DISCIPLINE FILOSOFICHE

AGENCY IN PROGRESS.
THE ETHICS, EVOLUTION, AND PSYCHOLOGY OF MORAL CHANGE

Tutor: Prof. Massimo Reichlin
Co-Tutor: Dott.ssa Sarah Songhorian

Tesi di DOTTORATO DI RICERCA di Federico Bina
matr. 015517
Ciclo di dottorato XXXV
SSD: M-FIL/03

Anno Accademico 2021/2022

RELEASE OF PHD THESIS

I, the undersigned, BINA FEDERICO
Registration number 015517

Born in Milan on January 7, 1995

Author of the PhD Thesis titled "Agency in progress: The ethics, evolution, and psychology of moral change"

AUTHORIZE the public release of the thesis.

Reproduction of the thesis in whole or in part is forbidden.

January 31, 2023

DECLARATION

This thesis has been composed by myself and has not been used in any previous application for a degree. All the results presented here were obtained by myself, except for:

1) Chapters 8 and 9 have been re-adapted from a manuscript titled "Individual Moral Progress: A Virtue-based Approach", co-authored with Maria Silvia Vaccarezza and Matilde Liberti (University of Genova), Sarah Songhorian and Massimo Reichlin (Vita-Salute San Raffaele University). The paper is under-review at the time of submission of this thesis.

2) Section 11.3 has been re-adapted from "Moral Progress: *Just* a Matter of Behavior?", *Teoria*, 42(2), 175-186, co-authored with Sarah Songhorian, Francesca Guma, and Massimo Reichlin (Vita-Salute San Raffaele University).

All sources of information are acknowledged by means of reference.

Acknowledgements

This work and several other accomplishments and opportunities in the past years would not have been possible without the contributions of several remarkable individuals.

First, I am deeply grateful to my advisors and mentors, Massimo Reichlin and Sarah Songhorian, for their kindness, patience, and guidance over these years. I thank them for the numerous discussions we had regarding the central topics of this work, but most of all, for the freedom, opportunities, responsibilities, and encouragement they have provided me. They are both amazing philosophers, great professionals, and dear friends. Working with them has always been a pleasure. If I can become even half as good as they are in this field, I will consider it a great success.

Thanks to Josh Greene for welcoming me into the Greene Lab for nearly a year. Josh has been an exceptional advisor and made me feel at home from the very first day. His ability to bridge different theoretical perspectives and to discuss several issues with his incredible multidisciplinary competence is unparalleled. I am truly grateful for his generosity, encouragement, and insightful suggestions for this work.

Thanks to Silvana D'Ottone and Lucius Caviola for their friendship, for the enjoyable experiences lived together and for valuable discussions for this work and on several other moral and existential issues. Thanks to Casper Gelderblom for being – for the second time – the best roommate anyone could ask for, and for the comraderie, great moments, and enlightening conversations we had in Cambridge. A special thanks to Davide Battisti, for his irreplaceable insights and reflections on moral and ethical issues. Without him, my understanding of these subjects would be far less clear. Thanks to Allen Buchanan, Matilde Liberti, Stefano Pinzan, and Hanno Sauer for important discussions and suggestions about central aspects of this work. A special thanks to Maria Silvia Vaccarezza, for her faith in me and our work together. I look forward to continuing this collaboration in the very near future. I am also grateful to all the members of the Greene Lab and the Moral Psychology Lab at Harvard University, as well as the members of the PROGRESS group at Utrecht University, for providing numerous opportunities for learning and discussion that have greatly enriched my understanding of many issues and this work.

I would like to extend my gratitude to many other fantastic persons who have had a more or less direct impact on this work through their conversations, ideas, and memorable experiences shared together. Their contributions have made it this journey more pleasant, and a more complete and fulfilling experience. Thanks to Giorgia Adorno, Vera Elizabeth Allen, Giacomo Arrigo, Luca Ausili, Gabriele Beretta, Livia Bresciani, Laura Burkhardt, Elisa Canu, Stefano Cardini, Dario Cecchini, Francesca Cesarano, Emanuela Ceva, Michel Croce, Michael Dale, Wim De Neys, Francesca Guma, Sally Haslanger, Michael Klenk, Victor Kumar, Charlie Kurth, Sergio Filippo Magni, Carlo Martini, Maurizio Mascitti, Matteo Motterlini, Vittorio Emanuele Parsi, Rik Peels, Francesca Pongiglione, Fausto Raschioni, Paul Rehren, Alessio Salviato, Eleonora Severini, Simona Tiribelli, Alessandro Volpe, Alessandro Volpi, Lea Ypi. Those I have inadvertently forgotten, I thank too.

Thanks to Lisa for her love and enthusiasm. Thanks to the PDC and εύρηκα, my two second families, and to Antonella, Marcello, Mauro, and Michela for being the best actual family that one could ask for. A special thanks to my grandma Pinu, for being an endless source of love and care. Her incredible flexibility in understanding and adapting to a changing world fills me with optimism every day. Finally, thanks to my brother Andrea, constant source of inspiration with his exceptional curiosity and openness to explore new worlds, cultures, and experiences, and to my mother Monica for granting me the invaluable gift of the freedom to choose my own path without any limitations. If I can speculate about the value of freedom in this work, it is thanks to you.

In fond memory of
Antonello (1952 - 2019)
a true moral exemplar

Abstract

In this work, I seek to address the complex topic of moral progress from some underexplored theoretical angles, aiming to offer an original perspective on issues that recently gained renewed attention by combining approaches and data from moral philosophy, cognitive science, history, evolutionary and social theory.

In the first part of the work, I explore various fundamental philosophical issues arising from the evaluative core of the inquiry about moral progress. After a conceptual and comparative analysis, I emphasize the need for normative and axiological ethical reflection in moral progress theory, also by responding to a recent 'non-ideal' objection to this approach. I defend an *agency-based* theory of moral progress, according to which referring to the value of agency and autonomy – in terms of their increased exercise both (1) as a *result* and (2) as a *driver* of social and moral change – seems a particularly reliable proxy to justify progressive moral shifts, and to accommodate problems of alternative normative criteria and types of moral progress (e.g. increased well-being or social equality).

In the second part, I show how such a (normative) theory is naturalistically coherent and empirically sound. More specifically, I rebut what I call the 'hard-wiring' thesis, i.e., the idea that human moral and prosocial cognition are rigidly bound to be exclusive and myopic by the evolutionary history of our species, and that this constitutes a severe obstacle to significant agency-driven moral change. Against this pessimistic view, I show how socialization and several other ecological and epistemic conditions can favor considerable psychological shifts which allow for significant increase in people's agency, ability for open-ended practical reasoning, and prosocial behavior. This, in turn, facilitates and stabilizes emancipative and inclusivist shifts both in individual values and social institutions.

In the third part, I explore the idea of an improvement of the moral capacities of individuals. I discuss the limits of available models in the literature and defend a naturalistic, empirically-informed, and action-guiding virtue-based account, emphasizing the importance of developing – apparently paradoxical – stable habits and traits of mental flexibility. Finally, I consider some of the limits of conceiving individual moral improvement in terms of skills acquisition, analogously to the development of expertise in non-moral domains, by also examining recent research on the psychology of moral learning and decision-making. I conclude by defending a procedural approach for assessing improvements and/or different levels of reliability in moral judgment and decision-making.

Abstract – Italiano

In questo lavoro affronto il complesso tema del progresso morale da alcune angolazioni teoriche relativamente poco esplorate nel dibattito contemporaneo, con l'obiettivo di offrire una prospettiva originale su questioni che di recente hanno guadagnato rinnovata attenzione combinando approcci e dati provenienti da diverse discipline (filosofia morale, scienze cognitive, evoluzionismo, teoria sociale).

Nella prima parte del lavoro discuto alcune questioni filosofiche fondamentali legate alla natura intrinsecamente valutativa dell'indagine sul progresso morale. Dopo un'analisi concettuale e comparativa dell'idea di progresso morale, difendo la necessità della riflessione etico-normativa e assiologica nell'indagine su questo tema, anche rispondendo a una recente obiezione 'non-ideale' a questo approccio. Propongo poi una teoria del progresso morale basata sull'*agency,* secondo la quale il riferimento al valore dell'agency e dell'autonomia – nei termini di un loro maggior e miglior esercizio (1) come *risultato* e (2) come *motore* del cambiamento sociale e morale – appare un criterio affidabile e coerente per la giustificazione delle valutazioni morali 'storiche'. Questa prospettiva, inoltre, risulta in grado di spiegare e risolvere alcuni dei principali limiti di altri criteri normativi-valutativi o "tipi" di progresso morale (e.g. incrementi di benessere o di eguaglianza sociale).

Nella seconda parte mostro come tale teoria (normativa) sia coerente con numerose evidenze empiriche e con un approccio naturalistico allo studio della morale. In particolare, critico la tesi – piuttosto diffusa nel dibattito contemporaneo – secondo cui la psicologia umana sarebbe rigidamente destinata a rimanere tribale, esclusiva e poco lungimirante per via della storia evolutiva della nostra specie, e che ciò costituirebbe un ostacolo per un miglioramento delle capacità decisionali e motivazionali degli individui. Contro questa visione pessimistica, discuto numerose evidenze che mostrano che diverse condizioni socio-ecologiche ed epistemiche possono favorire notevoli cambiamenti psicologici, contribuendo a aumento significativo dell'autonomia, delle capacità decisionali e delle disposizioni prosociali delle persone. Questi cambiamenti, di conseguenza, favoriscono e stabilizzano cambiamenti emancipativi e inclusivi tanto nei valori morali individuali quanto nelle istituzioni morali e sociali.

Nella terza parte discuto infine l'idea di un miglioramento delle capacità morali degli individui. Discuto i limiti di alcuni modelli presenti in letteratura e difendo un approccio naturalistico, empiricamente-informato e *action-guiding* basato sulle virtù che enfatizza l'importanza di sviluppare alcuni tratti stabili di "flessibilità mentale". Infine, considero alcuni dei limiti di intendere un miglioramento morale individuale meramente nei termini dell'acquisizione di competenze o abilità (*skills*), analogamente a quanto avviene in altri domini (non-morali), discutendo alcuni dati recenti sulla psicologia dell'apprendimento e delle decisioni morali. Concludo difendendo un approccio procedurale per la valutazione di miglioramenti e/o diversi livelli di affidabilità nei giudizi e nelle decisioni morali.

Contents

Cause-and-effect assumes history marches forward, but history is not an army. It is a crab scuttling sideways, a drip of soft water wearing away stone, an earthquake breaking centuries of tension. Sometimes one person inspires a movement, or her words do decades later; sometimes a few passionate people change the world; sometimes they start a mass movement and millions do; sometimes those millions are stirred by the same outrage or the same ideal, and change comes upon us like a change of weather. All that these transformations have in common is that they begin in the imagination, in hope. To hope is to gamble. It's to bet on the future, on your desires, on the possibility that an open heart and uncertainty is better than gloom and safety. To hope is dangerous, and yet it is the opposite of fear, for to live is to risk.

I say all this because hope is not like a lottery ticket you can sit on the sofa and clutch, feeling lucky. I say it because hope is an ax you break down doors within an emergency; because hope should shove you out the door, because it will take everything you have to steer the future away from endless war, from the annihilation of the earth's treasures and the grinding down of the poor and marginal. Hope just means another world might be possible, not promised, not guaranteed. Hope calls for action…

Rebecca Solnit, *Hope in the Dark*

Introduction

In recent years, several scholars have promoted narratives suggesting that humanity's moral progress is constantly growing and almost certainly destined to continue to do so. As we will see from the first pages of this work, the idea of moral progress is a complex one, and several theoretical tools can be used to reflect on it. In this work, I emphasize the importance of ethical reflection in the justification of our judgments and moral intuitions about what counts as moral progress. Against ideas of moral progress as mere production of 'better' or 'more desirable' states of the world (e.g. increased well-being) and/or as depending on supra-individual forces which transcend individual thought, action, and responsibility, this work seeks to propose an agency-based view of the relationship between individual and social moral progress.

The central normative thesis put forward in the first part of this work is that instances of social change are morally progressive when they involve increases in agency and decisional autonomy. Specifically, changes are morally progressive i) if they produce or promote an increase in the agency capabilities of individuals and/or ii) if they result from improved exercise of these capabilities.

There are good reasons to believe that many human societies are improving in this sense. Nonetheless, progress is never linear nor perfect: it often takes steps forward and some backward, and sometimes instances of progress and regress coexist, making it difficult to draw

all-things-considered evaluations. As long as the ratio is positive, we can probably speak of improvement; but this balance is fragile, and much still needs to be done. What is almost certain is that progress has not been, and will likely never be, completely achieved.

Even the most liberal and progressive societies face enormous moral challenges nowadays. While data suggest that the world is improving 'overall' in many respects, many serious problems persist and even arise. And even when injustice, oppression, suffering and cruelty decrease (though always too slowly and not for everyone), sometimes we think and feel that we may – and perhaps should – do better in light of the resources we (especially WEIRD people) have at our disposal.

If the world progresses mostly for the richest, and inequalities, oppression, and suffering persist even when better scenarios could be possible for the most disadvantaged with minimum costs for the better off, it is not clear that 'the world' is progressing, even if the well-being or average health of the worst off slightly increases. Contemporary societies continue to present enormous structural inequalities. The patriarchy continues to reign and women to be subordinated. Almost one in four women in the United States reports being raped or a victim of sexual assault[1]. Racism and discrimination persist. Despite a quite significant decline in homophobia in several societies, transgender people continue to be excluded, stigmatized and discriminated against. The Russian invasion of Ukraine and the consequent humanitarian crisis, the civil war in Tigray, the Nagorno-Karabakh crises and several other conflicts all over the world made 2022 one of the worst years according to indices of peace and existential security since the '80s. Two-thirds of the wealth created on the planet since the beginning of the COVID-19 pandemic has gone to the richest 1% of the population, and for the first time in 25 years the condition of those living in extreme poverty has worsened.[2] It is now clear that the effects of anthropogenic climate change are real, recognized by the scientific community, and devastating especially for some of the poorest countries nowadays, as well as for future generations and entire ecosystems all over the world.

Several other instances of social and political instability worldwide suggest that much can and should be done to improve upon the current conditions, both locally and globally. In the past decade, ideological and religious views, nationalist, authoritarian, xenophobic, and anti-scientific sentiments have gained strength, with serious consequences on political institutions and individual freedom. By observing this I am not suggesting any kind of 'regression thesis': the world is getting better in several respects and there are reasons to be optimistic. But these

[1] CDC (2020).
[2] Christensen et al. (2023).

and other examples suggest that regressive shifts can always be at the door, and lack of vigilance and effort could put important moral gains at risk. This very latter claim might suggest a naïve and too optimistic emphasis on individual responsibility in historical processes over supra-individual and non-intentional dynamics. The present work aims to support this claim by arguing in favor of a more balanced methodological approach within moral and social change theory – recently, significantly biased towards the structuralist/anti-individualist side (see e.g. Madva 2016; Sauer 2019; 2023).

While the solution – or at least progressive mitigation – of some of the aforementioned problems may be located at the structural and institutional level, individual responsibility and action play a crucial role in these processes. Both in democracies and less democratic (or blatantly autocratic) contexts, the causal contribution of individual values and action to socio-political change is fundamental (Inglehart 2018; Welzel 2007; 2013), and the positive gains of institutional change in terms of protection and enhancement of individual freedom are one of the most valuable things to defend and promote. What we should be optimistic about, and invest resources on, is human agency: both as a driver of progressive change and as one of the most important values to defend and promote to achieve more progress and to avoid the risk of regress. Whether we are acting on the individual or structural-institutional level, we have reasons to believe in the fundamental changing potential of enhanced human agency and decisional autonomy rather than in the promises of some sort of teleological and/or supra-individual law, process or entity.

This work challenges 'externalist' and teleological views according to which moral progress happens outside of individual minds and – at this point in history – it is (almost) guaranteed by supra-individual dynamics rather than being shaped by psychological, epistemic, and value change – i.e. changes in people's minds and morality. While teleological views do not believe that moral progress is 100% certain in the future (it's just very likely), they are basically 100% sure that moral regress will not happen, since current social structures and institutions are strong enough to prevent it. My view takes the possibility of moral regress more seriously, and reflects on both the structural-institutional and psychological conditions that can predict or hinder different kinds of moral change (either progressive or regressive).

I also largely criticize the belief according to which recent scientific research suggests that human psychology is bound by our evolutionary history and by several biases that are almost impossible to eliminate. According to these views, any project of improvement of individual psychological capacities through ordinary means – like education, information, reasoning, intergroup contact and exchange, and other kinds of reform – is basically futile or ineffective

in addressing the most urgent contemporary moral mega-problems. Most of my discussion is dedicated to showing that these pessimistic views are empirically unsupported, and to offer reasons and evidence in favor of a viable and more desirable alternative.

A core thesis put forward throughout the chapters is the claim – neither new nor original, but quite controversial nowadays – that the moral improvement of individuals and that of societies and institutions are deeply intertwined. When moral and social change is not 'agency-based' in the sense introduced above, societal moral progress is more fragile, and circumstances are more susceptible to setbacks and regression. Improvement in the opportunities as well as in the decisional capacities of individuals is one of the key elements to stabilizing, and even improving, progressive moral gains.

A second important methodological claim needs to be made. This work takes very seriously the contribution that recent scientific research about human psychology, morality, and social institutions can offer to enrich our understanding of ethics, also in a normative sense. The social and cognitive sciences are increasingly providing us with a clearer understanding of several conditions and mechanisms involved in, and facilitating or hampering, moral change dynamics. Understanding and exploring these mechanisms can provide us with useful and effective tools to promote and stabilize our idea of progress and, sometimes, even to draw normative conclusions (Bina 2022; Greene 2014; 2017; Kumar 2017). Nonetheless, what improvements and deteriorations consist of is difficult to determine by science alone. In addition to its convinced naturalistic methodology, this work emphasizes the need to engage in normative ethical reasoning and value theory to reflect about the topic of moral progress. In this respect, the emphasis that I place on improving individual agency should not only be understood as a requirement and a goal of moral progress, but also as a fundamental, flexible meta-theoretical means to further improve the very idea that we have of it.

Part I. Towards an agency-based theory of moral progress

Introduction

What is moral progress? Many thinkers have tried to answer this question, recently as in older times, from several perspectives, with disparate methodologies, and drawing sometimes different, sometimes overlapping conclusions. This chapter is dedicated to an ethical analysis of the idea of moral progress, and to the presentation of a critical and original view within the contemporary debate. I do not aim to reconstruct the multifaceted debate on this concept, nor its long history: such a project would require much more space, and it would not allow me to sufficiently focus on some core issues I consider underexplored in the debate and literature on moral progress, and in need of further discussion and clarification.

The chapter develops as follows. In chapter (1) 'The idea of moral progress', I clarify some important differences distinguishing this concept from other, related ones, and introduce an important distinction between two possible understandings of the concept – namely, moral progress as *morally desirable change* and as *progress in morality* (1.1). I then outline a minimal theoretical framework including four criteria that I believe any theory of moral progress should respect (1.2). In (2) 'The normative-evaluative core', I emphasize the need for normative and axiological ethical reflection in moral progress theory, also by responding to a recent 'non-ideal' objection to this approach (2.1). In (3) 'Agency-based moral progress', I expose my original view on the issue, setting the ground for a 'dual' agency-based theory of moral

progress. According to this view, referring to the values of agency and autonomy – in terms of their increased exercise as a result (3.1) or as a driver (3.2) of socio-moral change – is a particularly reliable proxy to justify progressive shifts and to accommodate problems of alternative criteria for moral progress.

1. The idea of moral progress

1. Conceptual and comparative analysis

According to a minimal conception, "Moral progress occurs when a subsequent state of affairs is better than a preceding one, or when right acts become increasingly prevalent" (Jamieson 2002, 318). This definition, however, seems far too minimal and vague, and asks for further conceptual specification. 'Better' how? What 'right' acts? And what justifies considering them as such?

Before presenting a more structured framework including four essential elements that I believe any theory of moral progress should contain, allow me to introduce a few indirect, comparative considerations about what the concept of moral progress is *not*, and how it differs from related concepts. First, moral progress is different from 'mere' moral change. I return to this point several times over the course of this work, but suffice it to notice this: moral change can be conceived of as a system of historical processes concerning shifts in moral institutions, practices, beliefs, emotions, theories, and so forth – observed and analyzed with a descriptive, rather than normative (or prescriptive) attitude. Theories of moral change aim at describing and explaining these processes; on the contrary, the idea of moral progress is an evaluative concept, which expresses a morally positive *evaluation* of instances of change.[3]

Second, despite the two concepts being closely related, the idea of moral progress is also different from the idea of moral development.[4] Technically, the latter refers to the ontogenetic process by which individuals develop morally relevant cognitive and behavioral traits. Classic

[3] As we will see in a moment, moral progress can be the evaluation of instances of both moral and non-strictly moral change, and here resides one of the main sources of controversies and misunderstanding in the contemporary debate. I am not taking a stance here on whether this evaluation consists of non-epistemic attitudes of approval or whether it tracks objective moral facts. For alternative legitimate positions on this issue see e.g. Hopster (2020), Huemer (2016), Luco (2019), Rorty (1999).
[4] On this point, see chapter 9 of this work and Schinkel & de Ruyter (2017).

research in developmental psychology has suggested that 'normal' moral development follows relatively rigid genetic stages – each specifically characterized in terms of, e.g., empathic and/or reasoning abilities –, typically moving from self-centered and egocentric to more empathic, controlled, imaginative, universalist, care-based perspectives (Hoffman 2000; Kohlberg 1981, 1984; Jubilee Centre for Character and Virtues 2022), despite significant disagreement among theories and relevant individual and cultural differences.

Although moral development plays an important role in the dynamics of moral progress, two main aspects differentiate the two: a) moral development should be understood, like moral change, as a process that can be observed, understood and described in non-evaluative terms; b) moral development refers to the ontogenetic evolution of single individuals' morally relevant cognitive and behavioral traits (though of course unfolding in communities through social learning and experiences) from early infancy to adulthood, while moral progress (typically) refers to broader historical and societal moral shifts.[5]

Partly analogous considerations can be made for the concept of moral *learning*. The concept of moral learning is also non-evaluative, and it differs from moral development for its reference to more abstract cognitive and/or computational processes of value acquisition and representation that are not necessarily limited to the early moral, social and cognitive development of humans or other animals. Moral learning processes also occur later in life, and some of their computational underpinnings can be modeled by drawing on formal methods developed in other domains, such as reinforcement learning, Bayesian inference, and other machine learning techniques (Cushman, Kumar & Railton 2017).

Before getting to a meta-theoretical framework for moral progress, a final important consideration is in order about the meaning of 'moral' in the aforementioned concepts.[6] On the one hand, as just stated, in the cases of moral change, moral development, and moral learning, the word 'moral' has – or at least should have – a descriptive, non-evaluative meaning: all these concepts typically refer to historical processes regardless of their being considered good or bad,

[5] As it will become clearer throughout the text, these are closely intertwined with the possibility of less pre-set and rigid individual moral transformations occurring during maturity. Even in this case, however, moral development would still describe and explain certain psychological and social dynamics independently of their being judged morally desirable or not. Such evaluative conclusions must be drawn, at least partially, on different bases. This does not imply that a descriptive psychological understanding of both early-'standard' and future, wider and more flexible/open-ended kinds of moral development cannot have relevant implications for a theory of moral progress. Still, this relevance is only 'indirect'. That is, descriptive understanding of how certain kinds of moral development actually occur would need to be combined with independent, philosophical discussion of what counts as progressive. On such an idea of the 'indirect' normative relevance of empirical data, see e.g. Greene (2014).

[6] On the definition of morality and on normative and descriptive approaches to this problem, see Frankena (1967/70), Gert & Gert (2020).

desirable or not.[7] Specifically, the concepts of moral change, development, and learning are not evaluative in a 'positive' sense: we can easily think of good and bad, virtuous and vicious, desirable or not, normal and pathological (etc.) examples of these processes. In all these cases, the term 'moral' simply refers to their involving changes in cognitive and behavioral traits, institutions (etc.) that are related to our generally accepted understanding of what morality is as a social phenomenon, or that are 'morally relevant' in a very broad and non-evaluative sense.

On the other hand, the concept of moral progress is intrinsically evaluative in a positive sense. But where does its evaluative element reside? This issue has been a source of disagreement among scholars in the recent debate on moral progress. I suggest there can be two main possible readings at stake here: one could attach the main evaluative element of the concept of moral progress either in the adjective 'moral' or in the concept of 'progress'. Let us briefly see how.

i) Moral progress as 'morally desirable change'. Those who defend this view tend to attach a positive value to the term 'moral'. When speaking of moral progress, they implicitly rely on a normative conception of morality and what is moral. Hence, according to this former view, the idea of moral progress refers to historical processes:

a) whose main subject of change and whose primary progressive element(s) can also be not *intrinsically* related to morality;

b) which are *also* 'moral' in the sense of being desirable or welcome from a moral point of view.

Think, for example, of techno-scientific or economic progress. These kinds of historical changes are not directly related to morality: first, they do not necessarily involve the exercise of, or change in, moral agency, reasoning, beliefs, norms, understanding, sentiments or motivation. Second, the criterion by which we evaluate their 'progressiveness' is not necessarily a moral one: it can be related to how well certain artifacts perform their function, to the fact that they make it easier to do something, to their implications on well-being, etc. (Buchanan & Powell, 48-53; Kumar & Campbell, 2022, 177; Sauer 2023). According to this view, the concept of moral progress can be applied 'widely' to a very broad number of types

[7] As widely discussed in Part II, defining 'moral' and 'morality' is a particularly complex task. However, while people might disagree about what morality is in descriptive terms (as a social phenomenon, institution, set of practices, etc.) this does not imply that this disagreement is *moral* or *evaluative* in nature, i.e. disagreement about what is good or bad, right or wrong. Morality and related concepts can still be understood as a set of real phenomena which exist independently of our evaluations, so that disagreement about what morality descriptively is – as a social institution or in our language – is not an instance of *moral* disagreement. Notice that this has nothing to do with moral realism, which affirms the existence of mind- or stance-independent moral facts that are *intrinsically normative*, and not the simple existence of morality or 'moral' phenomena as real phenomena that exist independently of our evaluative moral attitudes. More on this in Part III.

and cases of social change if their social implications are simply 'welcome', or desirable from a moral point of view, i.e. according to a very broad and pluralistic set of values and normative principles (Sauer 2023; Sauer et al. 2021). For instance, according to this understanding the idea of moral progress can be applied to situations in which institutions, social or economic systems produce better consequences in terms of subjective well-being, or if they become more equitable. According to this first possible meaning, the evaluative core of the concept of 'progress' can be independent of genuinely moral considerations (and grounded, for example, on prudential ones), while the use of the term 'moral' is used in a positive evaluative sense, similarly to its use in the expression 'a moral person'.

ii) Moral progress as 'progress in morality'. The second possibility consists of locating the main evaluative element of the concept of moral progress in the term 'progress', and to conceive of the word 'moral' with the same descriptive attitude we (should) have towards concepts such as moral change, moral development, and moral learning; that is, to refer to processes and changes more directly related to morality (e.g. shifts in moral institutions, norms, sentiments, etc.). According to this use of the concept, the idea of moral progress is closer to that of 'progress in morality', with a specific focus on the psychological and epistemic aspects – such as beliefs, reasoning and understanding, emotions, motivations, and so forth – involved in complex social institutions and systems of interaction and cooperation that permeate and shape the lives and identities of individuals and groups, and which also reflect on themselves producing theoretical systems (see Kumar & Campbell 2022, 179).

A few contributions to the recent debate on moral progress have reflected on this important but often neglected distinction, first emphasized by Buchanan & Powell (2018)[8]. Both Buchanan & Powell (2018) and Kumar & Campbell (2022) – two of the most relevant contributions to the debate – consider this distinction fundamental to understanding the concept of moral progress, and both conceive types of the second kind – progress in morality, moral capacities, even 'moral minds' – as a privileged kind of moral progress, if not the most important and the only 'genuine' or 'proper' one (Buchanan & Powell 2018, 63; Kumar & Campbell 2017, 179). Nonetheless, it must be highlighted that none of these authors provides a convincing justification for why such a distinction should be considered important, except for its undeniable intuitive appeal (Buchanan & Powell, 51).

Such a lack of justification has been recently emphasized by Hanno Sauer (2023, 3.6). Sauer highlights several inconsistencies entailed by accepting a sharp distinction between the two

[8] In slightly different terms (see section 3.2 below).

aforementioned meanings of moral progress – that he phrases in terms of 'broad' vs 'narrow' –, concluding on its fundamental inadequacy and uselessness (see also Kitcher 2017, 53, for a similar though less articulated view). One of the aims of Chapter 3 is to add more critical insights to this discussion, trying to understand whether this distinction can be justified from a normative-evaluative standpoint. Before doing that, however, let us consider some more general, basic elements that any sound theory of moral progress should include.

2. A meta-theoretical framework

In what follows, I provide a meta-theoretical framework including a few minimal criteria that any reasonable theory of moral progress should satisfy. This framework is meta-theoretical (or methodological) in the sense that it does not pose substantive constraints on what should count as moral progress, but rather on how the issue should be addressed by any theoretical project.

I suggest that any account of moral progress should include the following four theoretical elements:

1) an *evaluative-normative standard* to assess when, whether, and why trends or instances of moral-social change are morally progressive. Such a standard should work as a reliable proxy to state when, whether and why a state of affairs, institution, character (or something else) *x* at t_1 is morally 'better' compared to a previous configuration of *x* at t_0 (intra-comparison) or compared to another entity *y* (inter-comparison) either in the same period or in a different one (Rønnow-Rasmussen 2017);[9]

2) a *value theory* that justifies the normative-evaluative standard on a 'different level' (i.e. provides at least partial independent justification for it), and specifies the relations between intrinsic and extrinsic values. Value theory can be conceived of as "an abstract structure that specifies a set of fundamental values that, when conjoined with propositions about particular people, societies and so forth, implies a nested hierarchy of less fundamental values" (Jamieson 2002, 327);[10]

[9] Several scholars suggested that such a standard does not need to be monistic, and that it could (and should) rather be conceived as plural: "if there is only one normative standard by which to assess ethical gains – increases in aggregate welfare, say – then either there has been progress on that one metric or there hasn't been. With multiple measures of progress, the situation is more complicated: perhaps a society has managed to improve along one axis, but not without deteriorating along another" (Sauer 2023, 3.7). I discuss this issue in greater detail below.

[10] Jamieson calls "more fundamental values 'deep values', and less fundamental values 'shallow values'. Values can stand in relations of 'deeper' and 'shallower' with respect to each other" (Ibid.). As he adds, while this lexicon "may sound foundationalist [...] this way of representing a theory of value is consistent with alternative models, such as coherentist ones" (327-328, footnote 29). For a similar view, see also Brink (1989, chapters 5 and 8).

3) a set of reliable *assessment tools* to evaluate whether (and/or to what extent) the change or trend under evaluation is an instantiation of the evaluative-normative standard of reference, as well as whether promoting or realizing extrinsic values actually promotes or realizes intrinsic value(s);

4) a realistic, *descriptive theory of moral change* (non-evaluative) stating how moral shifts occurred in the past, and how they might be reliably promoted, stabilized, or avoided.

Together, these elements provide a minimal but sufficiently robust framework for understanding and assessing moral progress.

First, the *evaluative-normative standard* is crucial for stating whether, when and why a state of affairs, institution, character (etc.) is morally better than others (in particular, previous configurations of it). This standard should serve as the normative reference for determining whether an instance of change is morally progressive, and as a clear and consistent – though fallible – criterion to morally evaluate different states of affairs, institutions, or persons (inter-comparison) or the same object of evaluation in different moments in time (intra-comparison).

As we will see more in detail in the next section, however, despite significant overlapping on certain moral principles or norms, competing accounts endorsing or sympathizing with different normative and/or value theories may disagree about the 'moral progressiveness' of specific instances of social or personal change. Different ethical theories prioritize different values and goals, and may therefore come to different conclusions about the moral progressiveness of certain instances of change. Moreover, while some classic and very general normative ethical principles – such as increase in population welfare (cf. Evans 2017), gains in social equality, and moral inclusion can certainly be appropriate criteria to assess the progressiveness of moral shifts (see Kitcher 2011; Luco 2019; Sauer 2023, chapter 5; Scanlon 2018; Singer 1981/2011), in the next sections I will emphasize that the most discussed proxies (or types) of moral progress in the literature present important limitations, which can be explained and corrected by attributing a prior role to the respect for, or increase of, *agency* and *autonomy* in the evaluation of historical moral shifts.

As I will argue below (Chapter 3), such a reference to the value of agency and autonomy provides:

• A coherent and unifying (though not reductionist, monistic, nor foundationalist) justification for more 'shallow' normative-evaluative standards/proxies – such as equality and well-being – with which to judge instances of moral change as progressive, and to shed light on their main limitations;

• A reliable criterion to distinguish between different degrees of 'worth' in instances of moral progress, correcting too coarse-grained distinctions between 'social' and 'moral', 'wide' and 'narrow' moral progress (or 'desirable social change' and 'moral progress strictly speaking');

• A reliable criterion to distinguish between more and less *stable* instances of moral progress;

• The best criterion to guarantee the greatest possible normative-evaluative flexibility and open-endedness while overcoming the theoretical and practical limitations resulting from the excessive fragmentation and lack of normative-evaluative guidelines in the typologies of moral progress currently available on the market.

Several relevant aspects of items 3 and 4 of the meta-theoretical framework outlined above will be discussed in Part II and III, though an exhaustive discussion of them (especially point 3 – 'assessment tools') would require much more space, specific scientific competencies, and even different theoretical aims – mostly because of their essential empirical nature. Nonetheless, reliable assessment tools are fundamental in a theory of moral progress, since without them it would be hard to assess the actual progressiveness of moral shifts in light of given normative-evaluative standards. For example, to evaluate the moral progress of a society in terms of increased personal autonomy or agency, we would need to determine how these constructs are operationalized and measured. Several measures can be used, including quantitative measures of political and economic freedom, statistical analysis of qualitative data (such as surveys of individuals' perceptions of their own autonomy and agency), comparative historical analysis, case studies, interviews, and so forth. Once the constructs are operationalized, several statistical tools can be used to assess, for instance, whether a society presents higher levels of individual autonomy and agency than before, as well as correlations and/or causal relations between measures of agency, equality, well-being, and/or more inclusivist and democratic institutions (Welzel 2013)[11].

Finally, a reliable descriptive theory of moral change is necessary to understand how moral shifts actually occur (both desirable and undesirable, progressive and regressive, as well as those characterized by greater normative-evaluative uncertainty). Without reliable scientific explanations of how moral change did (or did not) occur in the past, it would be hard to develop

[11] As Sauer notices, such a view "involves very strong commitments to the possibility of moral comparisons between cultures: [...] societies in which children are beaten, most women disenfranchised, most men oppressed, and which are joyless, impoverished hellholes (think of various theocracies or authoritarian regimes today or throughout history) are less morally developed than some existing alternatives." (Sauer 2023, 53).

effective strategies for avoiding regressive shifts and/or promoting further improvements in the present and future, however conceived (Kumar & Campbell 2022; Kitcher 2021).

Some of the most paradigmatic and less controversial instances of moral progress, such as the abolition of slavery, the emancipation of women, reduction in racial and ethnic discrimination and prejudice, the gradual expansion of human rights, the condemnation of aggressive war, greater freedom of expression, increasing concern for animal treatment, future generations and ecosystems (etc.) occurred thanks to a combination of social movements, shifts in public opinion, theoretical reflection, technological and economic change, favorable ecological conditions, and more. A realistic descriptive theory of moral change should take into account the role and relative weight of these factors in specific cases by referring to reliable historical data and analytic tools.

A couple of final general considerations on the idea of moral progress are in order. First, as virtually any contributor to the recent debate on moral progress has stressed, the idea of moral progress has no need to rely on the postulate of the existence of a final, perfect, or ideal end-state to progress towards. As Sauer notices, "We are perfectly capable of making comparative assessments of relative improvement even in the absence of an anticipated end state of perfection" (Sauer 2023, 1.3). Consider a famous analogy suggested by Amartya Sen:

> If a theory of justice is to guide reasoned choice of policies, strategies or institutions, then the identification of fully just social arrangements is neither necessary nor sufficient. To illustrate, if we are trying to choose between a Picasso and a Dalí, it is of no help to invoke a diagnosis (even if such a transcendental diagnosis could be made) that the ideal picture in the world is the Mona Lisa. That may be interesting to hear, but it is neither here nor there in the choice between a Dalí and a Picasso. Indeed, it is not at all necessary to talk about what may be the greatest or most perfect picture in the world, to choose between the two alternatives that we are facing. Nor is it sufficient, or indeed of any particular help, to know that the Mona Lisa is the most perfect picture in the world when the choice is actually between a Dalí and a Picasso (Sen 2009, 15).

Nonetheless, I want to emphasize that this should not lead us to renounce reflecting about the normative-evaluative core of the idea of moral progress altogether. If we want to understand what makes instances of moral and social change morally progressive, denying the need for a justified normative-evaluative standard – however imperfect – would be a mistake. As I will argue below, such a standard must not necessarily be understood as 'ideal', nor as 'teleological'[12]. On the contrary, I suggest conceiving of it in negative, flexible, and open-

[12] Here I use the term 'teleological' in its evaluative-normative sense according to which the value of an instance of change should be measured according to its approaching a goal fixed in advance ('progress to') in contrast to an idea of improvement

ended terms, rather than as inherently postulating a clear, positive, determined fixed content (more on this below). According to this view, moral improvements should be understood, more convincingly, in terms of "a moral betterment relative to the status quo, where this does not entail that there is some endpoint against which improvement is to be gauged" (Buchanan & Powell 2018, 46). To put it differently, moral progress should be mostly understood as emancipation *from*, rather than as improvements *towards*. As Rahel Jaeggi observed by commenting Kitcher's recent thoughts on this issue,

> Equal consideration, inclusion, or democracy (in a Deweyan society that 'gives the oppressed a steady voice') is not a reified goal, not something we aim at and might gain as a trophy at the end of our journey but a mode of operation, a process (Jaeggi 2021, 135; see also Kitcher 2011; 2021).

A second point on which many scholars seem to widely agree is that judgments of global, all-thing-considered, and long-term historical moral progress should be avoided. While in the case of more local and short-term comparisons it is easier to identify causal connections between causes and effects, in the case of global evaluations too many variables are usually at play (Buchanan & Powell 2018; Jamieson 2002; Kitcher 2011, 242; Kumar & Campbell 2022, 181-184; Sauer 2023, 3.4).

Third, and very related, morally progressive and regressive shifts can co-occur at the same time (Sauer 2023, 3.9). For example, the increase in antibiotics administration can be considered both positive and negative, since while antibiotics have greatly improved the ability to treat bacterial infections, their overuse and misuse can lead to the development of antibiotic-resistant bacteria, making it more difficult to effectively treat infections in the future, and leading to the destruction of beneficial bacteria in the body, with very negative impacts on overall health.[13] Perhaps a more relevant and paradigmatic type of ambivalence concerns unequal increase in economic well-being. While this may appear as morally progressive for those who benefit from it (Evans 2017; McCloskey 2010, 26; Sauer 2023, 5.1), inequality in wealth typically also creates or exacerbates differences in autonomy and social power between people and societies (Bornschier 2002; Piketty & Saez 2003; 2014). In what follows, I will

conceived of as 'progress from'. The agency-based view I defend here is primarily non-teleological in this specific sense. In this I am very sympathetic with Kitcher's pragmatic account of moral progress (2021), though I prefer conceiving of progressive shifts in terms of *emancipation* rather than in terms of *problem-solving* (as Kitcher does). To be clear, I am not referring here to 'teleology' in the sense according to which there are some laws or supra-individual mechanisms that necessarily (or very likely) will lead humanity towards progress – a thesis defended by Sauer (2023).

[13] Easterbrook (2018) mentions the case of progress in the construction of weapons systems: while weapons like missiles and bombs have become more precise and deadlier over time, they have also become much safer to use (155).

offer some further insights to partly address the problem of evaluating ambivalent cases like these.

2. The normative-evaluative core

What makes moral change progressive – or social change morally progressive? When and why can we justifiably say that a state of things, institution, or character *x* at t_1 is better than its previous configuration of properties at t_0 with respect to morality? As introduced above, moral progress is an evaluative concept; one of the core claims I make in this chapter is that accounts and judgments of moral progress should be explicit about their normative-evaluative core (i.e., about the moral premises, principles, or intuitions they rely on).

This demands, first, reference to more 'superficial' or 'shallow' normative-evaluative standards that can work as proxies to assess progressive moral shifts, and which can, *prima facie*, even justify why they are progressive (both socio-moral shifts that already occurred and those that one believe should be promoted). Second, accounts and judgments of moral progress should also be clear about their 'deeper' or more 'fundamental' evaluative core (cf. Jamieson 2002, 327-328). That is: What is it that matters morally in social-moral shifts that makes it reasonable to deem them progressive? And what justifies the appropriateness and inappropriateness of more superficial normative standards as proxies for moral progress? This project has been, I believe, surprisingly underestimated by recent accounts of moral progress, despite its being, in my view, one of the most fundamental inquiries in which any theory of moral progress should engage. To make a few (oversimplified) examples,

> we can imagine a broadly consequentialist conception of moral progress conceived as the trajectory toward a human community with a greater distribution of intrinsic goods (welfare,

> pleasure, etc.). Alternatively, we can imagine a deontological conception of moral progress as the trajectory toward a community where agents more reliably discharge their duties, or a character-based conception of moral progress as the advancement toward a world populated with progressively virtuous characters. Of course, these are just some of the ways of conceiving moral progress; moral progress could consist in a wide range of increases in the relevant right-making and/or good-making properties (Evans 2017, 76).

For instance, Evans suggests a 'non-evaluative' proxy property for moral progress: population welfare. His thesis is that an increase in population welfare correlates with moral progress, so it should count as a proxy property to identify progressive moral shifts. The only thing Evans leaves out in presenting this correlation, however, is a quite important one – giving an even vague hint of what moral progress may be.

With the exception of a very few others, like Singer (1981/2022) or Luco (2019), most of the contributors to the contemporary debate seem to reject this approach (i.e. justifying moral progress judgments by engaging in normative-ethical and axiological reflection), preferring the development of more pluralist, multi-parametric indexes, descriptive typologies, and lists of paradigmatic instances of moral progress (Buchanan & Powell 2018; Jamieson 2002; Sauer 2023)[14]. These operations are certainly all very useful, and contribute to an extraordinary improvement in our understanding both of the dynamics of moral change and of the main theoretical problems involved in the idea of moral progress.

However, while such indices, typologies, and lists of uncontroversial examples of moral improvement can be helpful to orient us in historical moral evaluations – as well as in thinking and perhaps even designing future moral changes – they mostly enrich our conceptual understanding and descriptive knowledge of the dynamics of moral change. They can only moderately help us to point out which instances of change can be considered progressive (e.g. among novel and 'non-paradigmatic' ones) and, especially, why. In other words, they only provide very 'shallow' and coarse-grained tools to evaluate socio-moral shifts, often providing no 'deeper' moral justification of *why* those types or instances of moral or social change can be considered progressive. Several scholars have justified their choice not to engage in this

[14] The most paradigmatic cases of moral progress on which virtually everyone (at least in the philosophical literature) agrees are: the abolition and repudiation of slavery; the reduction of discrimination and prejudice based on ethnicity and race; the emancipation of women; reduction in the stigmatization and increased acceptance of gay people; the decline and condemnation of aggressive war, colonialism, apartheid, exploitation, and violence towards Indigenous people; abolition of the most cruel forms of punishment; the extension of political participation rights; greater freedom of expression and from religious persecution (see Buchanan & Powell 2018, 47-48; Kumar & Campbell 2022, 181). A more controversial case is the better treatment of nonhuman animals – considered by some an instance of progress (Buchanan & Powell 2018, 47), but by several others one of the clearest instances of moral regress (Huemer 2019; Kumar & Campbell 2022, 181; Sauer 2023, 2.3), because of the increasing number of non-human animals suffering from human causes (Fitzgerald 2008; Ritchie et al. 2017).

kind of ethical inquiry on different bases. In what follows, I will consider one of the most radical among them.

1. A 'non-ideal' challenge

One of the most radical objections against the idea that we should justify our evaluative judgments about socio-moral change via normative ethical reasoning or value theory has been recently raised and defended by Victor Kumar and Richmond Campbell (2022). In their important book on evolution and moral progress, Kumar and Campbell argue that such a demand for justification and for a "general ethical theory that explains why some changes are progressive and why others are regressive" (189) is basically an overly skeptical move, which relies on an old-fashioned – and incorrect – way of doing ethics. According to Kumar and Campbell, traditional ethical theories understood as generalizations that are supposed to justify more specific moral evaluations, are "lofty idealizations that lie beyond the limits of human knowledge"[15] (190), and such an "ethical code is, to say the least, very hard to come by" (ibid.). In fact, they add, "a universal ethical theory is much more controversial than what is being asked to justify" (197).[16]

It looks like Kumar and Campbell are the skeptics here, in a quasi-fideistic way. Since human limited rational capacities cannot grasp the moral truth – nor any other kind of sufficiently reliable, higher level of justification for moral judgments – we should stop inquiring and rely on some uncontroversial truths about a few instances of moral progress. In their view, these truths are not revealed by God or any other transcendent entity, but by rational evolutionary processes that vindicate them:[17]

> morality is the product of Darwinian processes that are much smarter than any individual, smarter indeed than any group of individuals engaged in collaborative reasoning [...]

[15] As anticipated above, and as it will become clearer below in this chapter and in Part III, however, acknowledging the need for engaging in normative and axiological ethical theorizing in moral progress theory does not necessarily imply committing to ideal moral theories or principles.

[16] As I will argue more in detail in Part III, I believe that certain moral 'truths' – not necessarily in a robust realist sense – can be justified, even though they are more general than specific judgments about particular cases (see e.g. Brink 1989; Sidgwick 1907/1981; Huemer 2017). Also, some specific judgments are certainly less morally controversial than others and we can claim, without many problems, to have sufficient intersubjectively justified beliefs about their correctness. But in other cases, the more specific we get – e.g. when we have to translate a principle into a norm, or norms into specific judgments – the more we encounter disagreement. Applying consistency reasoning and reflective equilibrium between judgments, norms, and principles about which there are different levels of reflective agreement and confidence could help us find commonalities between different ethical theories; from there, we can generalize at higher levels of abstraction (in a fallibilist, always revisable way) and apply the provisional principles – even just procedural ones – to new and/or more controversial social challenges and ethical problems.

[17] On the idea of 'moral vindication' see Kumar (2017).

> Darwinian processes have crafted a much more worthwhile system than any that humans are capable of inventing on their own. [...] It may seem as though you can build your morality from scratch, but this is an illusion (191).

This is certainly true, but we can evaluate our received judgments in light of other considerations and standards, revising them even radically (Brink 1989; 2014; Greene 2017; Singer 2005). There are good reasons to claim that human agency and reasoning abilities have been able to produce fairly 'smart' stuff – as the authors themselves acknowledge in their book – and that is not entirely explainable in Darwinian, selectionist terms (Buchanan & Powell 2015; Singer 2005; 1981/2011).

Kumar and Campbell correctly remark that "Moral conclusions cannot be drawn from purely factual premises", but this seems exactly what they are doing, and not only in one, but in two very problematic ways. First, they seem to violate their own warning not to trespass the fact/value distinction by claiming that evolutionary processes are 'smart' and biased towards progress (196; for a more systematic defense of a similar teleological view, see Sauer 2023). Such a view could be acceptable if they provided independent ethical reasons to ground the normative-evaluative core of their idea of progress. In that way, showing that evolutionary processes progressively promote it or realize it (though always imperfectly and never entirely) would be a feasible theoretical strategy. But by avoiding suggesting, specifying, discussing, or justifying any normative-evaluative element – there is of course no need to provide *the* only possible, universally correct ethical theory – their theory of moral progress remains a very sophisticated theory of moral change, but not a fully justified theory of moral *progress*.

Second, Kumar and Campbell claim to adopt a 'non-ideal' approach to moral progress theory, as well as to philosophical ethics more broadly. This strategy consists in starting from relatively uncontroversial moral judgments and paradigmatic instances of moral progress – e.g., that chattel slavery or racial subordination are wrong, and their abolition is good and just – to build ethical theories (or maybe just to inform our future experiments in living, since they appear extremely critical about the overall utility of any 'traditional' theoretical project)[18]. However, by emphasizing that "Clear cases of moral progress (and moral regress) are the most secure starting place for ethics, much more so than [...] universal ethical theories [...]" (192),

[18] A partly similar view has been defended in a recent paper by Lea Ypi (in preparation), where she suggests a personal reading of Kant's view of moral progress (also significantly filtered by Hegel's reading of Kant). In line with Kumar and Campbell – and perhaps less faithfully to Kant – Ypi suggests a radically opposite view to the perspective I am offering here. According to Ypi, we need to refer to the progressive moral achievements that humanity realized so far in order to justify to the skeptic the normative authority of morality (and not vice versa). Unlike Kumar and Campbell, however, Ypi proposes a much clearer normative-evaluative criterion to assess the progressiveness of moral shifts – i.e., the progressive realization of the kingdom of ends on earth (which to me looks fairly 'ideal', and much more adherent to Kant's original view). I will return on this point below in this chapter.

Kumar and Campbell seem, again, to cross the very is/ought line they warn we should not pass. It is true – and I agree – that we cannot formulate our moral judgments and theories from scratch: we always need to start somewhere (Neurath 1921; Rawls 1980; Reichlin 2018). But if we *only* started from people's beliefs and evaluative attitudes about historical shifts and trends – what are commonly seen as clear instances of moral progress by most in a given population – without any other standard or independent support to evaluate the reliability of these judgments, this means (once again) to derive moral conclusions from factual premises, as any judgment of the form "x is good/right for population P" is, and to have a slightly too optimistic consideration of our ordinary moral judgments – an overconfidence that risks leading to conservative implications. To infer that something is good or right from the fact that we/many people/our ancestors consider it self-evidently good or right is notably one of the most problematic fallacies and conservative methodologies in ethics (Hare 1952, 41-44 and 81-93; 1963, 30-35; 1972; 1981, 17; Hume 1739/2000; Mackie 1977, 68 and 72; Nowell-Smith 1957, 35-38).[19]

Although certain cases or issues are morally less controversial than others, simply referring to self-evidence about the correctness of particular judgments, received opinion, or superficial agreement is usually ruled out from acceptable methods for moral justification (Brink 1989; Hare 1972; Harris 2012; Singer 1974; 2005; Scanlon 1998). This is quite relevant in the case of historical evaluations, since if history, evolution and progress really are (even only partly) under human control, it would be useful to develop intersubjectively justified tools or procedures for the collaborative assessment of new relevant historical changes and challenges. Let me emphasize that I am not rejecting Kumar and Campbell's methodology: I just find it incomplete.

Without an explicit reflection on the normative-evaluative core of our moral judgments, both the evaluation of past events and prospective strategies to promote progress and avoid regress seem to lack any meta-evaluative criterion beyond contextual evaluative attitudes and beliefs. Kumar and Campbell are optimistic that their non-ideal approach might also identify strategies to promote further progress in light of how it has been achieved in the past (2022, 193). Certainly, we can know what contributed to producing past instances of moral change, and Kumar and Campbell provide a very detailed analysis of some of these dynamics. But if we only know what has been achieved in the past and how, without understanding and/or agreeing about *why* certain shifts are instances of progress, it is hard to understand – even just

[19] As well as a potentially regressive threat for liberal democracies. To mention just a very recent example, see Giubilini (2022).

partially and of course fallibilistically – whether there are good reasons to consider both old and especially novel, different circumstances progressive or regressive.

Kumar and Campbell might have a response to this worry. In their view, their non-ideal, naturalistic methodology consists of vindicating moral shifts in light of their being rationally driven (195-198; see also Kumar 2017): moral change is progressive when it is rationally promoted. On this account, we just need to empirically assess when moral and social changes are driven by a rational effort; there is no need to refer to any traditional ethical theory to assess the progressiveness of socio-moral shifts. This strategy, however, presents some limitations. First, Kumar and Campbell are introducing here (without noticing or deliberately camouflaging?) a normative element – *rationality* – to evaluate moral shifts. But in this way, what they are doing comes very close to what they wanted to avoid. Rationality is a normative concept, and even a particularly complex and controversial one: it is notably difficult to formulate a theory of it – as they believe it is the case for other normative theories – both in general terms and in its relation to morality (what is rational, and what is morally rational?)[20].

Second, it appears challenging to evaluate whether and when the instances of moral progress they have in mind are actually rationally driven or not. How can we make such an assessment? Third, we can think of several cases that are 'rationally driven' in which it is not clear why they should be also considered instances of moral progress (e.g. social movements, public policies, wars, scientific and technological developments). Of course, this just tells us that rationality might not be a sufficient condition for moral progress, but it could still be a necessary one. However, it is not even clear that rationality *as a driver* is present in any case of moral progress (more on this below).

Finally, since it is not clear why rationality should be an intrinsic moral value – it might be, rather, an instrumental, or extrinsic one – Kumar and Campbell should justify their view by engaging, I suppose, in the traditional ethical philosophical method that they reject. In the next section, I suggest a partly sympathetic, but also partly different view. In my account, socio-moral change is progressive when it involves the exercise or increase of people's agency and autonomy – not only when it is rationally driven.

[20] I suppose Kumar and Campbell might respond here by referring to two core ideas – with which I agree more than I disagree: improvements in factual knowledge, and moral consistency reasoning (Campbell & Kumar 2012; 2013). Although these are certainly important elements that can drive progressive shifts, they both appear to me still insufficient criteria to assess whether instances of moral change are progressive or regressive. Being merely driven by greater factual knowledge – or increased consistency with it – is not enough to speak of the greater rationality of a moral action. Moral consistency reasoning, while being a fundamental driver for moral change, still lacks clear normative-evaluative criteria to justify moral shifts and conclusions beyond simple confidence in one's intuitions and judgments: it is not an intrinsically progressive device, leaving open the possibility to produce vicious consistency as well.

Note that my critiques should not be understood in a realist, anti-naturalist, or foundationalist way, nor as the thesis according to which whoever engages in moral progress assessments should necessarily endorse a specific traditional normative or value theory. What I am claiming is needed in any theory of moral progress is rather the necessity to refer to any reasonable normative-evaluative standard to – even tentatively – justify why we have reasons to believe that the events or trends we are considering are actually progressive. For instance, because we can assess, in a specific instance of social or moral change, the presence of (or increase in) certain values whose importance would not reasonably be denied by anybody involved. After presenting my working idea of agency-based moral progress, I will return to some of these criticisms, and suggest that my account can accommodate critiques of excessive idealization, determinate fixed content, epistemic arrogance, vagueness, and arbitrariness. In Part II, I will address criticisms of psychological impossibility and metaphysical queerness.

3. Agency-based moral progress

Recently, several scientists and intellectuals have suggested that evidence shows the world is becoming an increasingly better place in many, valuable respects (Cohen & Zenko 2019; Norberg 2020; Pinker 2018; Rosling 2019) – some specifically emphasizing a few of them, such as economic prosperity (Deaton 2013), or the decline of violence (Pinker 2011). Philosophers, however, have provided deeper inquiries into the moral core of humanity's recent cultural evolution. Are there instances of moral progress as well? Is there something like moral progress in history? Do these questions even make sense?

Some of the earliest theoretical reflections on the topic in the ethical-analytical tradition answered these questions by focusing on a few fundamental principles to account for their ideas of moral progress. For instance, Peter Singer (1981/2011) paradigmatically recalled Lecky's image of an expanding circle of moral concern – which Singer conceived in utilitarian terms (i.e. in terms of an extended benevolent consideration of preferences/interests based on sentience); Ruth Macklin (1977) emphasized the presence of increased 'humanity' and 'humaneness' in social and political institutions; Peter Railton (1986) suggested one of the first among several naturalistic, functionalist accounts of moral progress that flourished in more recent years – both in realist (Luco 2019) and anti-realist versions (Kitcher 2011); Michele

Moody-Adams (1999) mostly focused on moral progress as an improvement in the understanding of moral concepts.

On the contrary, in the recent works of Buchanan & Powell (2018) and Sauer (2023) we find much richer and detailed *typologies* of moral progress, that is, complex taxonomies including several different kinds of progressive moral change. Buchanan and Powell mostly avoid directly dealing with normative and axiological issues, except, in few cases, to criticize either utilitarian accounts of moral progress (such as Singer's and Railton's) or the very project of dealing with axiological issues and value theory, which in their view would be too epistemically arrogant, and lead to problematic consequences (see Buchanan & Powell 2018, chapter 3).

Quite differently, Sauer faces the normative-evaluative core of the problem more directly (though without framing its effort in these terms). Unlike Buchanan and Powell's, Sauer's typology is not just a "map": it also identifies and discusses some core normative-ethical reasons and problems for and against considering criteria such as well-being and equality privileged metrics for moral progress. What Sauer does not do, however, is try to delve more into the relations between these and other values and criteria in the context of historical moral evaluations.

One of the main reasons why these authors avoid dealing more directly and deeply with normative and axiological issues is due to their common acknowledgement of a fundamental and irresolvable pluralism in these territories. Therefore, according to both views, a theory of moral progress should be itself pluralist in at least three ways: i) it should acknowledge that "there is or may be a plurality of valid basic moral principles" (Buchanan & Powell 2018, 93), ii) it should not be reduced to just one type of moral progress, and iii) should characterize moral progress "in an open-ended, provisional, epistemically modest fashion, taking seriously the idea that there can be and should be progress in how moral progress itself is conceived" (ibid., 376; see also Sauer 2023, 3.10).

I fully agree with all these requirements, and in what follows I show that directly dealing with the normative-evaluative core of our understanding of moral progress and of paradigmatic progressive moral shifts does not necessarily violate them. Specifically, I will show that referring to the value of agency, autonomy, and freedom can make our understanding of progressive moral change more coherent, still without committing to a strong monistic or foundationalist view according to which such values would reductionistically ground, explain, or entirely justify any other possible kind of moral progress or value involved in it.

Such emphasis on agency, autonomy, and freedom – that, as we will see in a moment, can be framed in two main different ways – makes my account an 'agency-based' theory of moral progress. Similar views have been defended in the recent sociological (Welzel 2013) and philosophical discussion about more or less directly related topics. However, either they present significant differences – also because they are not directly concerned about the topic of moral progress (e.g. Brink 1989; Herman 1993; Raz 1986; Sen 1999) – or they lack sufficient convincing development and justification of this idea (Buchanan & Powell 2018, whose view, however, also differs from mine in several respects).

My proposal does not claim to be exhaustive nor conclusive on this issue, nor do I aim at developing a full-fledged, robust, and definitive agency-based theory of moral progress – a project which, if worth pursuing, will be continued in the years to come. What I more modestly aim to do is to suggest that there can be at least a few reasons to take this idea seriously.

The way I conceive of agency here – its exercise and its potential improvements – is mainly descriptive[21] and closely related to the idea of social and decisional autonomy in practical reasoning. What I mostly refer to is the (increased) ability to represent possible alternative goals and courses of action, causal relations between them and the reasons supporting them, together with the abilities to a) revise value representations in light of new information, and b) to act upon them independently from the influence of social norms and pressures, habits, biases, pre-reflexive and inflexible cognitive mechanisms (Asch 1956; Bina 2022; Cushman 2013; Dolan & Dayan 2013; Graybiel 2008; Greene 2017; Hacker-Wright 2015; Henrich 2020; Singer 2005; Welzel 2013)[22]. When I say that instances of moral change are progressive when they involve the exercise or improvements in agency and decisional autonomy, I have in mind the exercise of these capacities and/or their enhancement.[23] I want to emphasize that conceiving of agency and decisional autonomy in this way does not mean positing any ideal of perfection or end-state to look at, nor does it necessarily entail a commitment to any specific metaphysical

[21] In the philosophical discussion on agency and autonomy, it is common to distinguish between descriptive (D) and normative (N) conceptions of these ideas. In a nutshell, D conceptions typically refer to a set of features that have to be fulfilled to speak of different degrees of agency or autonomy, while N conceptions refer to the moral principle according to which agents should be treated as autonomous. While the two conceptions are closely related, here I will always refer to a D conception of agency and autonomy, unless specified otherwise.

[22] As self-aware and imaginative animals, humans can imagine and desire to live free from external constraints and oppression (Deci & Ryan 2000; Sen 1999). The opportunity to believe in this possibility, however, can change depending on social and environmental circumstances. When these beliefs and desires are shattered and frustrated, people are less able to exert control over their lives and their decisions (Ryan & Deci 2000; Baumeister et al. 2009). Following Welzel (2013), I conceive of agency as also strongly related with people's concern for equality of opportunities and social justice. As Welzel's research shows, emancipative processes and values lead people to develop lower intolerance to discrimination and injustice, and greater tolerance for norm deviations that pose no harm to other people's integrity. As empirical data show, for instance, as agency increases, homosexuality, the emancipation of women and of discriminated and marginalized groups gets more tolerated, and behaviors that violate other people's integrity get less tolerated.

[23] Importantly, these cases can also include improvements in *respect* for agency and autonomy, if this leads to actual increases or improvements of them.

view about free will and determinism. Agency and autonomy always come in degrees: they are not on/off conditions or capacities (Reich 2002, 93).

Now, there are two main senses in which we can conceive of agency and decisional autonomy as playing a prior role in understanding and justifying why certain instances of change can be held morally progressive: socio-moral change is progressive i) when it leads to an increase in agency capacities and their exercise, and/or ii) when it is driven by them.

1. Agency as result

First, socio-moral changes can be considered progressive when they produce an increase in agency and decisional autonomy. As Sauer correctly points out, while gains in social equality, well-being, and inclusiveness are important measures of improvement and are often involved in the most paradigmatic cases of moral progress, none of them, taken individually, can be considered a reliable proxy nor a prior normative-evaluative criterion to identify and justify progressive moral shifts (Sauer 2023, chapter 5)[24]. My core claim here is that focusing on the promotion/enhancement of agency and decisional autonomy can accommodate these problems. Referring to improvements of these kinds not only better clarifies and justifies what counts morally in the most paradigmatic instances of moral progress, but it is also able to clarify and justify why other – still very important – criteria such as equality, well-being, inclusivity, demoralization, better compliance, better understanding (etc.) sometimes seem to be fit, but sometimes do not seem to work that well as proxies for moral progress.

According to this first aspect of my agency-based view – increased agency as a 'result' or 'consequence'[25] – instances of moral and social change are progressive if they produce increases in agency and decisional autonomy conceived as freedom and emancipation from several constraints and sources of oppression and dependence – such as unjustified and oppressive social structures, norms and institutions; lack of and/or disproportionate distributions of resources and power; exclusivist and prejudicial attitudes and beliefs, biases, and more (for a detailed typology of exclusivity and inequality, see Kumar & Campbell 2022, chapters 9-10).

[24] For an attempt to justify the superiority of increase in population welfare as a proxy for moral progress, see Evans (2017). As I mentioned above and as it will become clearer in what follows, I believe Evans' strategy is unsuccessful.

[25] Notice that the fact that increase in agency can be a consequence of something by no means signifies that it is an 'end-state': on the contrary, precisely for its characteristics and for the kind of evaluations that we are considering – historical ones – the greater the increase, the more open-ended its own possible consequences will be. I will get back to this point below.

In his typology, Sauer (2023, chapter 5) identifies nine types of moral progress – gains in equality, increase in well-being, expansions of the moral circle, contractions of the moral circle, increase in liberty and autonomy, fewer bad norms (or proper demoralization), better good norms (or proper moralization), improved compliance with valid moral norms, increase in moral knowledge. Buchanan & Powell (2018, 54-60) suggest ten: better compliance, better moral concepts, better understanding of the virtues, better moral motivation, better moral reasoning, proper demoralization, proper moralization, better understanding of moral status, better understanding of the nature of morality, and better understanding of justice.

A detailed discussion of the reasons why each of these criteria can (and rightly should) be considered an important type of moral progress would require much more space, and important contributions to this discussion have already been provided in their typologies both by Sauer (2023, chapter 5) and Buchanan & Powell (2018, chapter 1). However, let me emphasize once again that these authors mainly abstract more general 'types' of moral progress from some of the less controversial and paradigmatic instances of moral progress reported above. To put it differently, these authors – and especially Buchanan and Powell's work – tell us that there are differences in the kinds of moral progress that occurred over history, and that these are the types of moral progress that can take place. By doing so, they mostly remain on a descriptive, taxonomic level, without engaging too much in discussing why the moral changes that occurred in those kinds of shifts are good, better, or more right than before.

In what follows, I suggest that each of the types of moral progress these authors identified can be held progressive *because* and *when* they lead to improvements in agency and decisional autonomy, and/or if they are driven by them (3.2). Moreover, referring to agency and decisional autonomy also allows us to explain why the other criteria listed above sometimes do *not* work that well as proxies for moral progress, and why adopting them as such would lead us astray.

Increase in equality and well-being. While a detailed discussion of each type of/criteria for moral progress would deserve more space, suffice it to notice, for the moment, that several influential contributors to the history of ethical and political thought over the past century have pointed out several good reasons why two of the most intuitive proxies for moral progress – increase in *equality* and *well-being* – should not be considered important *per se*, but only when and since they respect and/or they contribute to improve the opportunities and agency of individuals, and the capacities which allow and favor its exercise (Anderson 1999; Brink 1989; Dworkin 1984; Gewirth 1978; 1996; Nozick 1974; Nussbaum 2000; 2011; Rawls 1971; Sen 1985; 1999). First of all, in fact, both absolute and relative lack of goods and resources – e.g.

lack of primary goods and structural inequalities – significantly constrain the opportunities of individuals, not only materially and in terms of social power and freedom, but also for the negative effects these constraints have on the liberty of others.[26]

One might object to this view that if we accept a form of prioritarianism (hence excluding any possible leveling-down objection to egalitarianism), an improvement in the position of those who are worse off would constitute an instance of moral progress even if it does not increase their agency and the change is not agency-driven (see Fig. 2 below for a compatible case). However, data show that if the well-being of a population increases there is a high likelihood that the agency/autonomy of that population will also increase (Evans 2017; Sauer 2023; Welzel 2013), though it may still remain highly limited if the inequality with another (part of the) population is considerable. According to my agency-based account,

a) if the well-being of the worst-off increases, and their agency also increases, the change can be considered an instance of moral progress;

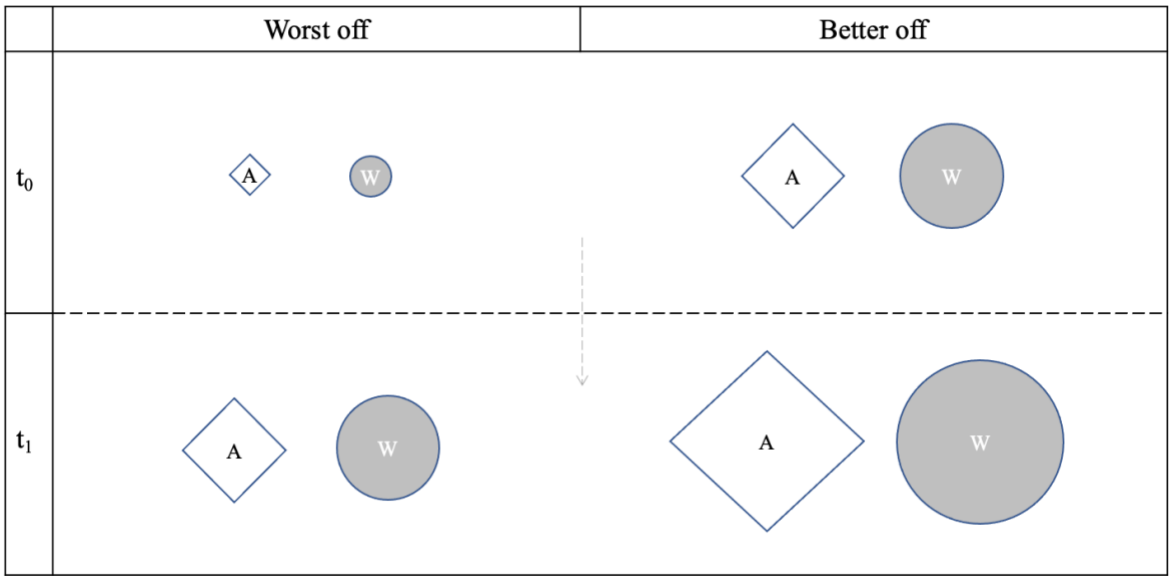
b) if the well-being of the worst-off increases because it is intentionally promoted by the better-off, even if the resulting increase in agency for the worst off is only moderate, that can be considered an instance of moral progress;

c) in the (maybe rare) case in which the well-being of the worst-off increases, but i) this does not lead to an increase in the agency/autonomy of the worst-off, *and* ii) it is not agency-driven (e.g., it happens by chance), while the state of the world is 'better' in some sense, according to my view there are no reasons to consider it an instance of moral progress (see Fig. 2).

Since these aspects are crucial to understand my agency-based account, below (Fig. 1-5) I consider a few more examples of societal change involving different combinations of increase in agency-as-result (A), presence of agency-as-driver (→) and their relation to (increases in) subjective well-being (W) and (in)equality to make clear how my accounts would evaluate them.

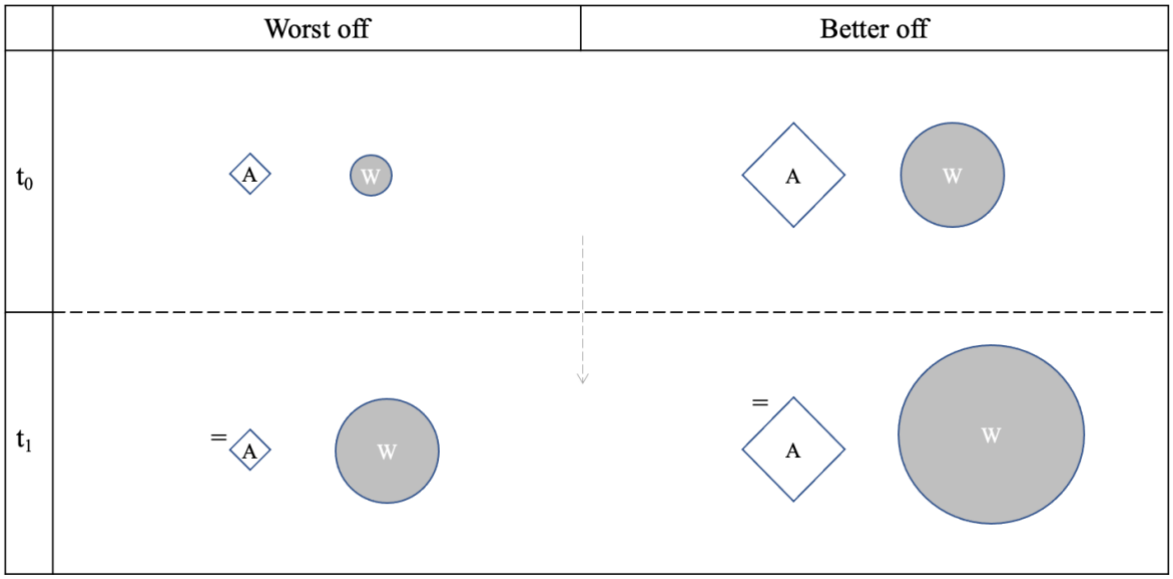
[26] I am grateful to Massimo Reichlin and Josh Greene for pushing me to clarify these points.

Fig. 1



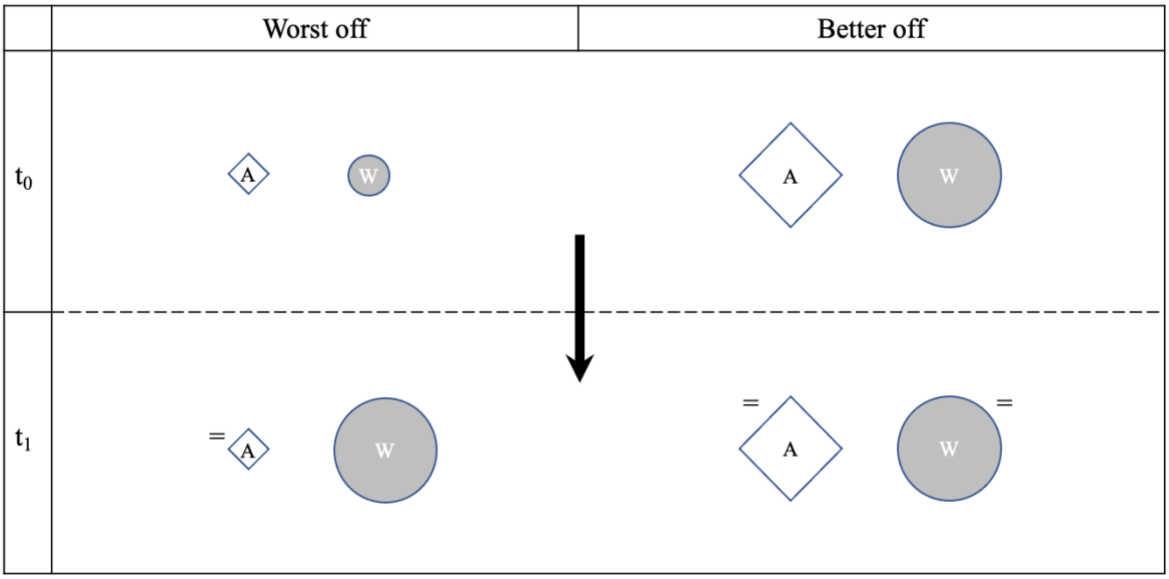
Case 1. At t_1, both the worst-off and the better-off of society S have an increased level of both agency and subjective well-being compared to t_0. The dashed arrow indicates unclear, weakly or non-agency-driven moral change. Moral progress: YES.

Fig. 2



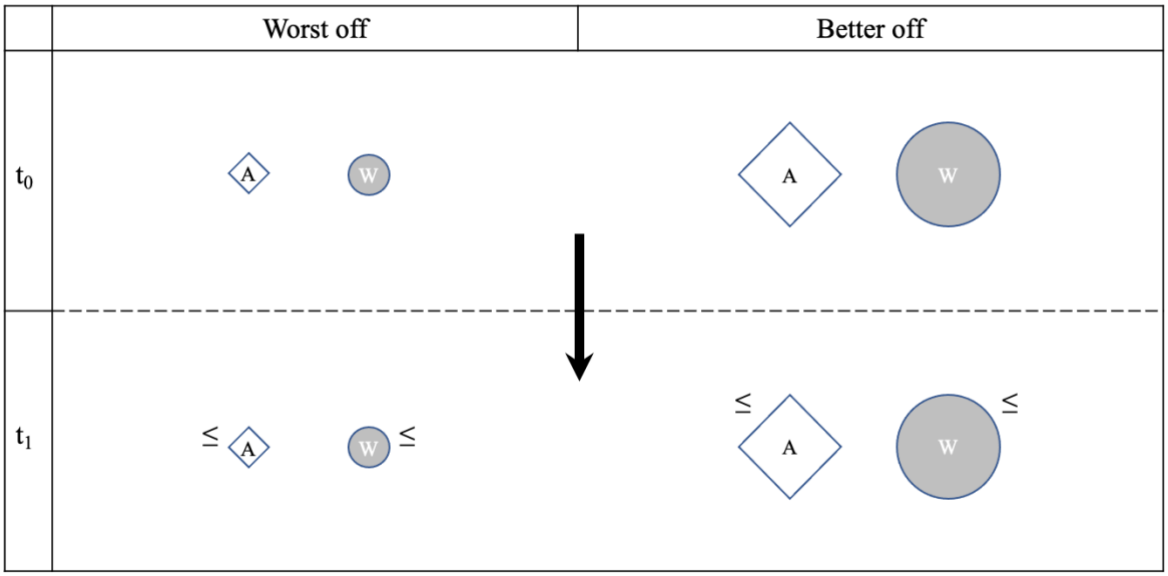
Case 2. At t_1, both the worst-off and the better-off of society S have an increased level of well-being but the same level of agency compared to t_0, and their increased level of subjective well-being is not clearly agency-driven (e.g. fortuitous or unclear). Moral progress: NO.

Fig. 3



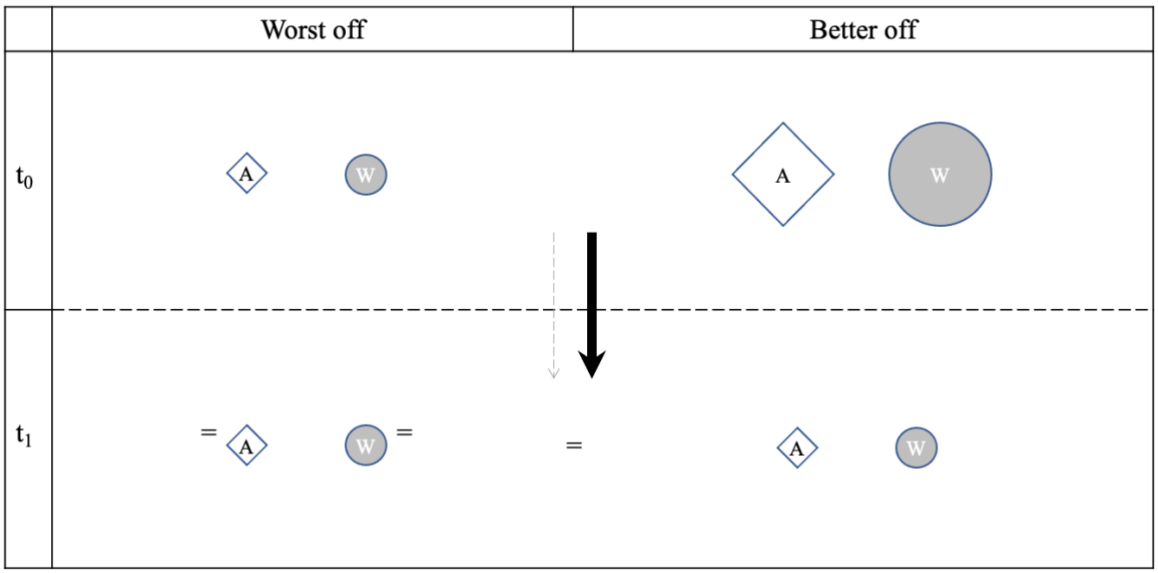
Case 3. At t_1, the only/most relevant change in society S compared to t_0 consists of a clearly agency-driven increase in the subjective well-being of the worst-off (no matter if produced by the worst-off, by the better-off, of by both). Moral progress: YES.

Fig. 4



Case 4. At t_1, both the worst-off and the better-off of society S present an inferior or equal level of both agency and well-being compared to t_0, after a clearly agency-driven effort to promote (or resist) change. Moral progress: NO.

Fig. 5



Case 5. At t_1, both the levels of agency and well-being of the better off of society S have decreased to the level of the worst-off, which has remained the same compared to t_0. Moral progress? NO (whether fortuitous or clearly agency-driven).

An old-fashioned consequentialist may still raise the following objection: What if increasing our agency/autonomy leads to a world in which we, as autonomous agents, end up making terrible decisions that make us miserable until the sun burns out? Is that good? Is that progress? By contrast, a world filled with blissfully happy creatures with very limited agency/autonomy, seems to the consequentialist like a pretty good world – certainly better than the high-autonomy, high-misery world.[27] I think that this objection does not hold for two main reasons.

First of all, the scenarios just outlined appear to me to be empirically implausible. It seems that the only reason someone might consider them empirically plausible – i.e., uncoupling agency/autonomy and happiness/well-being (hereafter *aa* and *hw*) so to conceive a world with very high levels of *aa* and very low *hw* (world A), and, conversely, a world with very high *hw* and very low *aa* (world B) – would be the endorsement of a subjectivist theory of *hw*, like a purely hedonistic or desire-satisfaction view. In fact, objective list theories (see Parfit 1984, Appendix I) typically include *aa* as necessary conditions for the realization of *hw* (Brink 1989; Nussbaum 2000; 2011; Sen 1985; 1999), so that, within an objectivist framework, both world A and B look basically impossible to imagine.

[27] Thanks to Josh Greene, Massimo Reichlin and Sarah Songhorian for making me clarify this point and for valuable discussion about it.

However, even if we agree to disentangle subjective *hw* and *aa* and we admit the possibility that an increase in *hw* does not necessarily correspond to a proportionate increase in *aa* (and vice versa) – intuitively, it seems plausible to imagine worlds with different, unbalanced levels of the two – it still seems implausible to have a scenario with an extremely high level of one and an extremely low level of the other. On the one hand, a large decrease in *hw* would almost necessarily bring a considerable decrease in *aa* (i.e. lack of resources constrain opportunities of action), so world A still appears very unlikely. Also, empirically, (subjective) *hw* are generally privileged enabling conditions for the development and increased exercise of *aa* (Inglehart 2018; Welzel 2013; more on this at the end of Part II), so that high *hw* would likely – though not necessarily – lead to at least a moderate increase in *aa*. On the other hand, the subjectivist seems committed to accepting that a reduction in *aa* does not necessarily produce a decrease in *hw* (otherwise, the theory would collapse into an objectivist one).

Hence, if we agree that even the subjectivist is committed to recognizing that only a moderate version of world A is empirically plausible, A would constitute an instance of regress only if/when a decrease in *hw* leads to a decrease in the *aa* of individuals: since if people decide autonomously to reduce *their own* level of *hw*, we typically would not consider it a problem (on the contrary, if agents would autonomously harm/violate others' *aa* we would, but see point 2 below). If our hypothetical world gains in *aa* but loses in *hw* in a way that prevents its population from developing or exercising further *aa*, that would not be an instance of moral progress. But if *aa* stays high and *hw* has been reduced autonomously and with the objective of increasing agency in the future (e.g.: people autonomously decide to reduce car driving to avoid the consequences of climate change, or decide to fight against an invader by sacrificing a lot of resources and *hw* investing in their future *aa*), that's certainly not a case of moral regress!

Let us now consider world B. A world with high *hw* and low *aa* would be positive just in the case that there are good probabilities for a further increase in *aa* (greater opportunities, lower costs in the development of prosocial dispositions and behaviors – see Part II, ch. xx; Buchanan 2020, xx). But in the case in which that does not happen – high subjective *hw* and low *aa* (both concurrent and subsequent of an increase in *hw*) – there are no reasons to treat this as a case of *moral* progress. Consider in this case Nozick's experience machine (Nozick 1974, 42-45): would we all be ok with the idea of having good (or better) experiences without being the authors of our choices? Or: would we all be ok with a world in which there is a non-violent and non-bodily harmful form of slavery, where enslaved people are treated kindly (or even more

kindly) by their oppressors, but they are still not free and both enslaved people and oppressors are happy(-er) with their conditions since they can see no alternatives?[28]

Or consider a world where ecosystems and animals are thriving, but humanity has gone extinct. Would that extreme case of world B (B^X) be better than world A, according to the non-consequentialist? I suspect that to answer that question negatively we need to accept a personal theory of value, according to which – *contra* Moore (1903/1993, 83-84) – only humans can be the source of moral value (see Brink 1989, 218-220). So it is not the simple capacity to feel pleasure or pain, or to have interests that has a moral significance, but rather the capacity for agency – the ability to distance oneself from, and to subject our habits, beliefs and actions to critical scrutiny and to act upon it that constitutes the source of moral value and normativity (Korsgaard 1996, 120-123). Of course, in B^X, nonhuman animals would still suffer, but there would be no problem with that; the moral problem with factory farming is not that animals suffer *period*, but that they suffer disproportionately because of human action when humans have good reasons not to do it and could act differently but do not.

If humanity were extinct, and there were a dramatic climate change causing a huge number of animals to die and suffer, there would be no reason to judge that event an instance of moral regress (not even of regress *tout court*), while it would still constitute a tremendous reduction in *hw*. Or think of a world B^N with very happy humans in great harmony with nature but deprived of many liberties that (some) humans nowadays have. Would it be a better world? Again, such a world would be quite hard to imagine, but I think that many of us would be inclined to say no (while this would certainly be an interesting topic for future empirical research).

A second answer to the objection concerns the conception of *aa* at stake. I already clarified above what I mean with these terms, but further discussion may be needed here. The idea of *aa* upon which my account is based should not be misinterpreted as the freedom to do whatever one wants, but rather as the ability to consider more and new different perspectives, goals, values, and ways of living; to consider the consequences of each of them and the coherence between the expected values of pursuing these options and the values or interests that one and others already have. Importantly, this includes the consideration of the perspectives, interests, and reasons of other subjects (Darwall 2009; Forst 2014; Railton 2017; Stueber 2017).

Recall that, above, I defined an increase in *aa* in the social sphere as the emancipation from limiting social habits, norms and pressures that typically discourage the possibility of thinking

[28] For some critical discussions of subjectivist theories of happiness/well-being, see e.g. Brink (1989), Colburn (2011), Nozick (1974).

otherwise and correlates with absolute duties to respect those norms, and to be loyal and more strongly bound to one's cultural identity and group members. At the same time, empirical research shows that greater individualism and social autonomy correlate with greater attention to the interests of others and with greater intolerance towards the violation of others' interests (see footnote 22 above and section 6.2 below – subsection on *market integration*). Thus, an increase in *aa* so understood is unlikely to lead to a drastic reduction in *hw* or in the *aa* of others, since one of its main features is precisely the actual ability (not only the intention) to take into consideration and to respect the perspectives, interests, reasons and autonomy of others.

Finally, remember that what is under debate here are historical moral evaluations, i.e. judgments comparing typically a previous and a subsequent state of affairs (see section 1.2 above). So the point is nor whether worlds A or B are good or bad in absolute terms, but whether they are better or worse compared to a previous configuration.

Inclusivity (or the expanding circle). While acknowledging its undoubtful importance in a complete typology and in the history of moral progress – according to Buchanan and Powell, progressive moral inclusivism is "possibly the most important type of moral progress" (2018, 63; for a similar claim see also Kumar & Campbell 2022, chapter 9) – understanding moral progress primarily a gradual expansion of the circle of moral concern presents some limitations. Recently, several scholars have stressed that Peter Singer's central idea that "morality is exclusively or almost exclusively concerned with promoting others' interests, and that moral progress consists in the move from considering the interests of a small group to considering the interests of ever larger groups [...] is far from capturing all of the moral evolution that we observe over human history" (Huemer 2017, 1998; see also Sauer 2019; 2023, 5.4).[29]

One of the main limitations of equating moral progress with moral inclusivity or expansiveness is that this appears unable to account for several other important instances of moral progress. As Sauer notices, "the elimination of taboos on, for instance, premarital sex, the decline in social punitiveness, or the abolition of dueling and footbinding cannot plausibly

[29] It has been pointed out to me that Singer's view of progress as expanding the circle is a statement about how progress tends to play out, not a strict criterion for what counts as progress. On the one hand, this confirms my claim that many influential accounts mostly describe how (desirable) moral change occurred and/or how it is still playing out, rather than providing one (or more) principle(s) to justify why that kind of change is progressive (or good, or desirable). One the other hand, if the expanding circle is correct as a descriptive thesis, we may still ask what makes it a good or progressive trend. The same old-fashioned consequentialist who pointed that out to me also told me that the basic utilitarian principle can provide that kind of justification, and I am inclined to agree. I do not see my agency-based view as incompatible with a (perhaps unorthodox) kind of consequentialism, though I am not sure that labeling it as 'utilitarian' would be fully correct. But such a discussion would lead the current one off-track. For a similar compatibilist view, however, see e.g. Brink (1989), Railton (1984).

be described as expansions of moral concern" (2023, 5.4).[30] I suggest that, once again, the wrongness of these practices and the progressiveness of their elimination appears justifiable, primarily, by referring to the constraints they pose to individual freedom and agency (and in a way which is able to account – unlike subjectivist views of well-being like desire-satisfaction theories – also for the infamous problem of adaptive preferences).[31]

Better morality for all!. Finally, both Buchanan-Powell and Sauer include in their typologies items such as "better moral concepts", "better moral norms", "proper moralization", "better compliance with valid moral norms" (and similar others). As anticipated above, however, these criteria are not very helpful if we have no meta-criterion whatsoever to judge when moral norms, concepts or reasoning are better or worse, valid or invalid, harmful or not, or when moralization or demoralization are proper or improper. All these types might well fit a superficial, formal, and descriptive 'map' of kinds of moral progress, but to say that moral progress includes the development of better moral norms and concepts does not sound like a great step forward. Is there any possibility of addressing this issue more deeply, perhaps by specifying what we mean by 'better' and why?

I do not seek to claim what makes moral concepts and norms good and bad, better or worse, (de)moralization proper or improper (etc.) in general. But I think that referring to improvements in agency (or in its exercise or respect) – as a *result* or *consequence* – can be, again, a promising route. Some of these types of moral progress – e.g. better moral norms, better understanding of justice and the virtues – can be judged as actually progressive when they result in greater respect or increase of the agency and decisional autonomy of people whose possibility to deviate from the most disparate constraints and forms of oppression (enslavement, social exclusion, discrimination, rigid social roles and identities, epistemic injustice, ignorance, existential threats, etc.) is/was otherwise restricted by social-moral norms, structures, and institutions. This does not mean that these instances of moral progress can be considered progressive only when this happens, and only because of these reasons: another fundamental reason why changes in concepts, motivation, and other types of moral change can be considered

[30] While considering it a favorite type of moral progress, Buchanan and Powell still notice that to reduce any kind or instance of moral progress to moral inclusivism would be a mistake: "This is true, for example, in relation to the moral reclassification of objects or entities that have no morally considerable interests of their own, such as sacred artifacts, non- sentient organisms, or abiotic features of the environment like rivers or mountains – at least when according such entities moral standing imposes unacceptable costs on beings that warrant moral regard. Fetishism, understood as the mistaken attribution of human or superhuman powers to nonconscious material objects, is an instance of "expanding the circle," but it is not moral progress; in some cases, it is a costly moral error" (2018, 63-64).

[31] See Buchanan & Powell (2017; 2018, 240-243).

morally better or progressive, in fact, is when they involve the exercise of moral agency (or its improvement) in practical deliberation as their *driver* (3.2).

A closely related consideration concerns 'better or improved compliance', an item included both in Sauer's and Buchanan-Powell's typologies (a point I will come back to in Part III). Compliance with moral norms or principles is not inherently morally right or good, since the norms and principles at stake can be unjustified: in these cases, improved compliance with 'invalid' moral norms can be an instance of regress, or (at least) a very strong obstacle to moral progress. If 'better or improved compliance' always implicitly refers to 'valid' or justified moral norms, the problem remains that of understanding and justifying when and why certain moral norms are better or more valid than others. But even in this case mere compliance with justified moral norms and principles may be a problematic criterion (or type of progress). A more critical acceptance or reflective endorsement of norms may be a better one, more closely related to the idea of *agency as a driver* (Korsgaard 1996) but also functional to the respect and/or promotion of increased *agency as a result*: understanding the reasons in support of a norm can make both its respect and justified exceptions easier; moreover, only if those reasons are sufficiently clear to oneself they can be intersubjectively communicated (Bina 2022; Scanlon 1998; Songhorian et al. 2022).

Similar considerations can be made for Sauer's 'improvement in moral knowledge'. Since it appears that we have no idea of whether an objective (mind- or stance-independent) moral reality exists, the only kind of moral knowledge we can reliably say to possess concerns intersubjectively justified principles, norms, and judgments (more on this in Part III). In my view, such an improvement can be traced back, once again, to improvements in agency, in the sense that principles are more justified when more potentially affected parties have the possibility to democratically express their voice and be their co-authors and co-legislators (Kitcher 2021; Ypi, manuscript).

But moral knowledge can also be conceived in a relativist, descriptive sense, so that we can imagine local improvements in what only certain people believe to be objective moral knowledge. Improvements in moral knowledge in this sense can be another strong obstacle to moral progress, since it can rigidly constrain agency in each of the relevant ways I consider here. Moral systems that claim to possess objective moral knowledge can constrain both the exercise of agency in practical deliberation (the possibility to think outside the system from within) and the respect and/or improvement of the freedom and decisional autonomy of others and/or in the future (see Buchanan & Powell 2018, chapter 3; Gaus 2016).

A further reason for this first kind of increase in agency and autonomy as a proxy for moral progress – and one of its main strengths compared, for example, to Buchanan & Powell's theory – is that it is able to account for progressive instances of social-moral change that occur fortuitously or as the result of causes that are mostly beyond human control (or not clearly attributable to human agency)[32]. According to my view, even unintended or only very moderately agency-driven socio-moral changes can be considered instances of moral progress if they produce an increase in people's agency and decisional autonomy – also because of the potential that such increases in agency may have for future agency-driven moral progress.[33] Therefore, a 'dual' agency-based account such as mine seems perfectly able to allow for

> societies to make progress even though the progressive change is unrelated to any policy they pursue. A small, poor, country is the only source for some substance; technological change generates a world-wide demand for the substance; the citizens become much richer, and their capabilities are considerably enhanced. Or an entrepreneur, intending only to make a profit, undertakes a venture with similar effects. Should we allow progress to be a matter of accident? I claim that we should. The pragmatic point of clarifying social progress is to enable future social change to proceed more "intelligently" (in Dewey's idiom). But that is compatible with — perhaps even dependent on recognizing that the bulk of past progress has been achieved blindly or accidentally (Kitcher 2017, 64).

Since here we are considering historical moral evaluations, increase in agency as a consequence of acts, events, or processes can (positively) affect further historical developments, which is where the second element of my account comes in.

2. Agency as driver

In the previous section, I suggested that a first aspect of my agency-based view consists in locating the value of progressive moral shifts in the respect and promotion of peoples' capacities for agency and decisional autonomy, and that this move seems able to justify both strengths and weaknesses of the competing normative criteria and types of moral progress considered by Buchanan and Powell and by Sauer in their recent accounts. I will now move to an even more controversial thesis, that is, that only changes that "come about through the

[32] According to Sauer, this is one of the weakest points of Buchanan and Powell's theory, and of excessively 'narrow' accounts of moral progress more generally (Sauer 2023, 3.6; see also Kitcher 2017).
[33] As already stated, it is important that such an increase does not create, maintain, or reinforce excessive inequality in agency and decisional-autonomy, since this would result in a significant lack of freedom for the more disadvantaged.

exercise of those capacities are instances of moral progress in the most full-bodied sense" (Buchanan & Powell 2018, 46).[34]

This view is based on the distinction, introduced above, between different ways in which one can understand the idea of moral progress. While above I considered only two of them – as Sauer (2023, 3.6) and Kumar & Campbell (2022, 177-180) do – Buchanan and Powell consider three main possible varieties:

> [1] moral progress in the most full-bodied sense is not simply change that is desirable from a moral point of view but also must involve the exercise of or improvements in the moral powers.
>
> [2] The second and weaker understanding allows changes that are improvements from a moral point of view to count as moral progress even if they came about through self-interested, prudential, or other nonmoral motivations (i.e., without the exercise of the moral powers or improvements of them). On the second understanding, Emperor Caracalla's extension of rights to a larger class of individuals would count as moral progress, but the reduction of disease due to a naturally mediated decline in parasites would not.[35]
>
> [3] The third and weakest understanding of moral progress would equate it with changes that are desirable from a moral point of view, without requiring that any human motivational capacities be involved. On the third understanding, the reduction of disease due to factors completely independent of human motivation and action would count as moral progress (Buchanan & Powell 2018, 51).

Buchanan and Powell flatly reject the third understanding as a justified kind of moral progress, and name type-1 cases 'moral progress in the robust sense' and type-2 simply 'moral progress'. However, as noted by Sauer, "we are given no reasons for this, other than that they happen to find it important" (2023, 3.6). Sauer acknowledges the intuitive appeal of this choice:

> What is the difference this distinction is trying to capture? Perhaps it can be explained in terms of the distinction between the wrong and the bad. Tsunamis are bad, but they are not wrong, presumably because they do not involve any human agency. High infant mortality is, provided that nothing can be done about it, merely very bad, but not morally wrong. It makes little sense to declare tsunamis impermissible, however devastating they may be. Social progress, then, reduces the bad; moral progress, narrowly speaking, mitigates the wrong (2023, 3.6).

[34] Buchanan and Powell have been the first to defend this idea in the contemporary debate, and their view still remains quite isolated. An only partly similar view – for some of the reasons discussed above – has been suggested by Kumar and Campbell: "When progress is moral, in our sense, the world improves not just as a by-product of other cultural developments (e.g. medicine), but through changes in the moral minds of human beings" (2022, 179). As we will widely see in what follows, several scholars in the contemporary debate – such as Macklin (1977), Sauer (2019; 2023) openly reject this claim.

[35] According to historians, the extension of the Roman citizenship promoted by Caracalla was a purely strategic move to curb independentist movements, get more taxes, and create bigger armies.

Nonetheless, Sauer also advances several objections to this distinction, which makes him "reject it, or at least deemphasize its importance", and conclude in favor of a wide definition of moral progress as mere morally welcomed social change (3.6). Against this view, in what follows I argue that there are reasons to consider this distinction important also by replying to some critiques moved by Sauer (2023), showing that they are not conclusive.

Remember that my agency-based view acknowledges that certain instances of social and moral change can be justifiably considered morally progressive even if they are not significantly agency-driven but, in the first place, when agency is improved as a *result* of the event, trend, or process under consideration. Hence, unlike Buchanan & Powell and in line with Sauer, I agree with drawing the distinction "within the concept of moral progress" (Sauer 2023, Introduction), without categorically excluding type 3 cases from my account. Nonetheless, the distinction between different senses and 'degrees of importance' of moral progress suggested by Buchanan & Powell remains important. Unlike its original proponents, however, my 'dual' agency-based theory is able to provide a preliminary justification and the normative-evaluative tools to assess differences of worth in progressive moral shifts (hopefully, further research and discussion will help to improve it in several respects).

As promised, let us now proceed to address some of the main critiques that have been (or may be) moved against this view. First, Sauer observes that

> wide and narrow progress cannot be meaningfully disentangled in real life: moral progress without social progress is empty. We would think that improvements in people's morality – their moral codes or beliefs – without any tangible social pay off would be hollow and worthless, even eerie and perverse. For instance, we would not recognize it as moral progress if everyone came to agree that slavery is abhorrent, with no discernible effect on the practice of slavery or the well-being and rights of the (previously) enslaved and their descendants. In fact, there would often be something particularly morally obscene about mere narrow moral progress without corresponding social gains (2023, 3.6).

I entirely agree: as I stressed in the previous section, the first of the two core criteria of my agency-based view requires that instances of social-moral change have a tangible effect in increasing people's agency in order to be considered morally progressive. As I argued above, these changes can be, and often are, gains in social equality and well-being, which subsequently lead to increase in human agency. If they do not lead to such an increase (e.g. gains in equality without increase in, or respect for, agency – see Fig. 5 above) it makes no sense to consider them instances of moral progress.

A second critique moved by Sauer against drawing a clear distinction between mere improvements in well-being and more intentional improvements in human morality (cf. Kumar & Campbell 2022, 179), is that

> improvements in well-being deserve to be classified as a form of moral progress because the most important of them – improvements in health, wealth, and safety – have been brought about by value-guided cooperative efforts to improve the human condition. They are the direct result of moral action (2023, 5.1)

and

> in almost all cases in the real world, even improvements that appear to be merely wide "social" progress will be in some sense due to human achievement rather than happenstance. Even diseases and poverty don't usually reduce themselves. Even the death toll of natural disasters is for the most part dependent on human cooperative accomplishments (3.6).

Consider the decrease in deaths by lightning. Not only, Sauer suggests, did these advancements increase people's freedom, but they are also agency-driven. As Pinker observes,

> what about the very archetype of an act of God? The projectile that Zeus hurled down from Olympus? The standard idiom for an unpredictable date with death? The literal bolt from the blue? […] [T]hanks for urbanization and to advances in weather prediction, safety, education, medical treatment, and electrical systems, there has been a thirty-seven-fold decline since the turn of the 20th century in the chance that an American will be killed by a bolt of lightning! (Pinker 2018, 189)

But this is false in many other circumstances. As Sauer himself notices on several occasions, many other desirable social gains are brought about by unintentional processes that are largely beyond human control (Sauer 2019; 2023; see also Acemoglu & Robinson 2012; Henrich 2015).[36]

And even if reduction of deaths by literal lightning – as other analogous cases – were moderately agency-driven, the degree of agency involved in or improved by this process (in the sense in which I characterized it above, and both as driver and as a result) is significantly lower than the agency involved and improved, say, in the case of the abolition of slavery or in

[36] For instance, Acemoglu & Robinson (2012, chapter 4) show how the Black Death in the 14th century had a big positive impact on the European economic, political and moral institutions in the years to come. Before then, the European economy was a highly unequal feudal system. But the great loss of lives the epidemics caused produced an unprecedented labor shortage, and wages rose up; this pushed Europe towards a different economy dependent on wage labor, who turned out to be much more efficient and productive. The bargaining power of workers was also positively affected by such a high demand for labor, and this contributed to the dismantling of the power of lords and, progressively, to more inclusive institutions – people started to have more opportunities, rights and liberties. Moreover, the Black Death killed more people in cities than in rural areas, balancing the geographical distribution of power, which also contributed to weakening the feudal system and led to more inclusive institutions. Thanks to Josh Greene for suggesting me this example.

women's emancipation. For instance, nobody started to deviate from oppressive moral norms in the case of the decline in lightning death, while this is exactly what happens in paradigmatic emancipatory processes. This is, I believe, what makes us judge the latter as more morally progressive than the former: the emancipatory force involved in them – both as a driver and as a result.

The point, once again, is not to see if there are reasons to hold Buchanan & Powell's distinction just as an interesting theoretical or explanatory tool, but also if there are reasons to hold it morally significant. The dual agency-based view seems able to account for that: improvements in agency and autonomous decision-making in paradigmatic instances of moral progress – such as women emancipation or the reduction of racial and ethnic discrimination – involve a much greater expression and improvement of agency and freedom *both as a driver and as result* than a reduction in deaths by lightning strikes. And this is precisely what makes them paradigmatic instances of moral progress.

What must be emphasized here is that there are not just cases of moral progress and cases that are not (or clearly distinguishable instances of social vs. moral progress). Moral value, or 'worth' often seems to come in degrees and – as in other kinds of moral evaluations – the difference in the degree of moral worth between two instances of change does not mean that one has full value and the other has none. Reduction in deaths by lightning strikes might have more moral worth than a reduction of diseases due to a natural decline in parasites, but less than reduction of the same diseases through a cooperative human effort, and even less than the end of a criminal military occupation after the efforts of a resistance movement.

This does not mean that the process that led to reduction in deaths by a bolt of lightning is flawed. Whatever increases moral worth – agency, reasons, motivations, virtues, costs – is just present to a different degree in the above cases. To have a lower degree of what enhances it does not necessarily mean to have zero. Some instances of social and moral change can be more progressive than others even if the former have been only moderately agency-driven and the latter are more clearly and significantly intended, designed and participated (such as, say, the abolition of slavery vs. the joint operations of international organizations and charities in response to localized natural catastrophes).

As Kumar and Campbell point out

> progress does not entail perfection. Nor does imperfection, even severe, entail lack of progress. To put this another way, things can get morally better without being good enough. For example, the legacies of chattel slavery and colonization are still painfully present across

> contemporary societies. Nonetheless, the world has seen moral improvement, even if only fragmentary, though their reduction (Kumar & Campbell 2022, 182).

Considering different combinations of the two main measures that I suggested – greater agency as result and greater agency as driver – might help us to assess different levels of worth involved in progressive moral shifts.

But why should agency-driven shifts be considered more worthy than shifts produced with much less significant involvement of human agency? First, agency-driven social and moral change usually involves an understanding of the situation and the values at stake and reflective interests to bring about change, which allows greater ability to take responsibility for beliefs and actions, and to justify them to others. Better abilities to understand morally salient issues, the values at stake and the reasons in favor and against certain courses of actions can facilitate the stability, transmission, and improvements of certain progressive shifts. This is more difficult in the case of accidental ones – even if they are the source of good consequences – where there might be no available reasons to justify them, nor future learning and transmission to realize them again in different contexts (for a similar point, see also Bina 2022; Kumar 2017; Kumar & Campbell 2022, 195-198; Sauer 2017; Songhorian et al. 2022).[37] As Sauer notes,

> many practices that used to be widespread but have meanwhile been abolished have become "beyond the pale". It seems inconceivable for, say, Denmark to decide to try slavery again, or roll back women's suffrage. Indeed, it is comical to even entertain these possibilities. Why is this? One explanation is that as societies reach certain levels of moral development and maturity, it becomes impossible for them to fall behind those levels (without disintegrating entirely). One way to conceive of "improvement" in moral understanding, beliefs, and motivation is through the lens of rationality. This means that moral understanding, beliefs, and motivations are considered to be improved when they are more rational and consistent with one another, and when they are grounded in evidence and reason rather than bias and emotion (Sauer 2023, 1.4).

But rationality alone might not be enough. Evidence and reason can make us perform very well even fairly unprogressive actions (or achieve unprogressive goals) if other kinds of resources, information, and possibilities are limited (see e.g. Hacker-Wright 2015). On the contrary, as mentioned above, when individuals have greater agency, their participation in collective decision-making processes is more likely to expand both their freedom and that of others. Moreover, exercising agency is an indicator of stability also because it requires the presence of

[37] I am grateful to Josh Greene and Massimo Reichlin for helpful discussions on this idea of cross-contextual stability.

certain relatively stable traits, and/or the respect for reliable epistemic and decision procedures (both in the case of individuals and institutions. More on this in Part III).

An already anticipated objection to my view could be that unequally distributed increases in the exercise of agency might be unjust and/or lead to undesirable consequences. Above, I emphasized the importance of gains in social equality to improve people's agency, since unjust social structures and hierarchies oppress people by constraining their freedom and opportunities (on each side, though of course in different ways). As I suggested when introducing my conception of agency, however, empirical research shows that this kind of freedom is strongly correlated to concern for equality of opportunity, social justice, and respect and tolerance for other people's freedom to deviate from social norms when no harm and limitation to others' freedom is involved (Welzel 2013).

Furthermore, one might also object that my project of finding a relatively clear and consistent normative-evaluative standard to assess old and promote future instances of moral progress is exposed to many critiques moved by Buchanan and Powell against some of the accounts they criticize (2018, chapters 2 and 3). Above all, it could be objected that, because of my narrow focus on a few core evaluative concepts, my agency-based view risks collapsing into what Buchanan & Powell call a *determinate fixed content* account, that is, a reductionist understanding of moral progress that arrogantly claims to have found the definitive criterion with which to assess moral progress also in the future, when our evaluative standards might even radically change from those that we consider justified nowadays.[38] According to them, in fact, "human beings have often (perhaps more often than not) been wrong about some aspects of morality and [...] there is no reason to believe that the sources of their errors have been eliminated" (Buchanan & Powell 2018, 92-93).

While I do understand the reasons behind this concern, an agency-based account appears to me fully able to accommodate this critique and avoid this problem. The emphasis on agency understood as the possibility to open-endedly and critically deviate from determinate normative standards precisely allows, and even requires, the greatest possible flexibility on the normative-evaluative level (Raz 1986). As correctly suggested by Buchanan & Powell (2018, chapter 3), a theory of moral progress should not specify a rigidly determinate value as an ideal to promote (see also Gaus 2016; Kumar & Campbell 2022, chapter 8), since this would not be sufficiently open to the possibility of revising our own theories and judgments of moral progress in the future. My proposal to focus on increases in autonomy and agency (also prospectively-

[38] According to Buchanan & Powell, in these cases "the substantive norms by which progress is gauged have three key features: they are (1) fixed or unchanging, (2) of determinate content, and (3) fully knowable here and now" (2018, 70-71).

teleologically, as a value to promote – see Herman 1993, 231; Wood 1999, 130 and 157-158), however, is probably the best possible way to guarantee this open-endedness requirement. Increases in agency (also as a function of increase in factual knowledge; Bina 2022; Greene 2017; Kumar & Campbell 2022, 197; Songhorian et al. 2022) are precisely what would allow moral agents to "engage in ongoing critical scrutiny of the norms they are currently adhering to" (Buchanan & Powell 2018, 85; Buchanan 2012), to continually re-evaluate and adjust normative standards and judgments in light of new information, and to consider diverse perspectives and engage in dialogue with them (Kitcher 2021).

If this is true, however, a further critique might be that such an emphasis on agency and freedom is *too* open-ended and that, ultimately, this would make my criterion vague and its implementation arbitrary. Which criteria (if any) should in fact regulate the exercise of agency and decisional autonomy in moral contexts, in practical interactions and deliberations? I believe that several theories have provided fairly convincing and radically non-consequentialist answers to this question, from the Kantian, contractualist, and critical theory tradition (Gewirth 1978; Habermas 1990; Rawls 1980; Scanlon 1998), from pragmatist perspectives (Kitcher 2011; 2021), virtue ethics and epistemology (Baehr 2011), and even in terms presented as potentially acceptable by any normative and metaethical perspective (Schaefer & Savulescu 2019). I will discuss this issue more extensively in Part III.

The evaluation of social and moral changes that are put forth and justified by moral reasons is a complex task, as it raises questions about the standard by which we can consider these reasons the most justified ones. My agency-based view is compatible with a procedural perspective where the justified reasons are those that would not be rejected by anybody in conditions where all potentially involved parties have the freedom to democratically express themselves (Kitcher 2021). This procedural, democratic decision-making approach cannot but be favored by, and favor, an increase in the agency and autonomy of individuals compatible with the highest possible degree of agency for others.

A final and more 'specific' consideration might concern the fact that an agency-based account such as the one outlined here is unable to account for the better consideration of and treatment of nonhuman animals as an instance of progress. However, both criteria of my approach can be sufficiently satisfied in the case of the increased moral concern for the condition of other nonhuman species. First, we can conceive of a weaker sense of my 'agency as result' criterion in terms of increased freedom for animals. Although freedom in this case has to be understood in a slightly different way, it can still account for the value of liberating nonhuman animals from several forms of constraints and oppression, and increasing their

chances of living a life as freely as possible according to their interests and capacities. Second, such an instance of change can be considered progressive when and because it is driven by the exercise of human agency. Greater concern for the condition and moral status of animals, as well as their better treatment, stems from an increase in humans' decisional autonomy (the possibility to think that we can treat them differently from exploitation) and agential capability to act accordingly to genuinely moral reasons (Shafer-Landau 2012). According to Sauer,

> Perhaps cosmopolitan attitudes and a concern for animal welfare are best captured by expansions of moral status. Perhaps, however, they should be described as the demoralization of tribal boundaries and species membership. The increasing number of entities enjoying moral status may well be a welcome by-product of more fundamental changes in what we regard as valid moral norms. (Sauer 2023, 5.3).

Perhaps, I add, they might be better understood in light of our increased liberty to choose what to eat and how to live, and that we can decide not to cause unjustified suffering to them because we have alternative options that are not too costly for us. Were these options reduced, and the interests of humans and non-human animals in conflict, for instance, one might think that human rights should be prioritized over those of animals, as many do believe even among animalists (consider the paradigmatic case of the ethics of experimentation)[39]. Of course suffering is bad, but it becomes morally wrong when it is unjustified, or when it can be avoided without sacrificing anything comparably bad.

As already mentioned, while I am aware that much still needs to be done to develop a full-fledged agency-based theory of moral progress, I hope I have been able to convey the idea that we may have at least a few reasons to take this view (or a part of it) seriously. This work will leave, for now, normative and evaluative questions behind, but the above ethical considerations justify the methodology that I will follow in the next chapters, characterized by an emphasis on a particular kind of moral change – psychological moral change. The priority that I will give to the analysis of individual-psychological value change and to improvements in moral decision-making abilities in the next chapters is thus justified by the theory that I defended in this chapter, although the content of the discussion (especially in Part II) remains in great part non-evaluative.

As I mentioned in my 'meta-theoretical framework' above, in fact, a realistic, descriptive theory of moral change is fundamental for any theory of moral progress, independently of any normative-evaluative framework of reference. As we will see in the next chapters, agency and

[39] See Kitcher (2015).

decisional autonomy require specific psychological capacities and socio-environmental circumstances to be exercised and enhanced. In particular, in Part II I will discuss the idea that human evolution might have considerably constrained the possibility of significant psychological change and, therefore, of robust agency-driven moral change. In the final chapters of Part II I will suggest how increased agency, and the peculiar human abilities for open-ended moral reasoning and normativity can be explained within a naturalistic but non-reductionist evolutionary framework.

Part II. Evolution and psychological moral change

Introduction

In this central part of the work, I discuss and criticize the idea that human moral and prosocial cognition are rigidly bounded by the evolutionary history of our species and that these constraints constitute serious obstacles to significant psychological moral change. Within recent scientific and philosophical debates, several scholars have defended this claim by relying on functionalist-adaptationist hypotheses about the natural evolution of moral cognition and institutions. Advocates of such 'hard-wiring' thesis state that selective pressures in small, close-knit pre-modern societies forged a tribal, exclusive, and myopic psychology that still affects and constrains human cognition and behavior, and that is relatively impervious to change – at least through 'ordinary' means of moral reform. In the next chapters, I address the empirical soundness of the hard-wiring thesis and provide several sources of evidence and arguments against it.

An important methodological caveat: this Part investigates the possibility of certain kinds of moral change, rather than speculating about their goodness or desirability. Therefore, the analysis carried out here aims to be mainly descriptive rather than evaluative; hence, a whole range of normative and meta-ethical issues related to the idea of moral *progress* will not be directly addressed in Part II. As discussed above, moral progress is an evaluative concept that implies positive ethical assessments of moral or societal change. Evaluative standards – and

moral ones in particular – are notoriously often controversial, and a defense (or critique) of normative and value theories grounding specific accounts of moral progress requires an (at least partly) independent ethical analysis, which is carried out in Parts I and III of this work. As already stressed in Part I, on a descriptive level the concept of moral change covers a whole range of historical processes and phenomena that can be studied empirically, suspending ethical judgments about their desirability. In these chapters, I will follow this non-evaluative approach.[40]

Notwithstanding, several ethical and political implications can of course be drawn from the non-evaluative issues considered here. It is no coincidence that recent empirical research in cognitive and social psychology keeps producing (more or less optimistic) normative reactions and strategies for psychological, moral, and social reform. However, recent discussion of these proposals has primarily focused on their moral acceptability rather than on the scientific validity of some of the empirical premises they rely on.

Consider, for instance, recent debates about *moral bio-enhancement* and *nudging*, two of the most influential projects claiming to rely on experimental evidence from the cognitive and behavioral sciences to steer people's behavior toward morally better directions (Persson & Savulescu 2008; 2012; Thaler & Sunstein 2007). Critics of these strategies have problematized their ethical assumptions and/or implications (Harris 2012; 2016; Rebonato 2012; Reichlin 2019; see also chapter 8.2 below), while much less attention has been devoted to assessing the scientific validity of some of their grounding empirical premises.[41] This latter task is what I set out to do in these chapters.

Empirically-informed strategies such as nudging and moral bio-enhancement aim to improve people's decisions and behavior claiming to rely on scientific evidence about the functioning of the human mind. Specifically, these projects emphasize that recent psychological research highlights the systematic influence of several biases and constraints in human cognition, decision-making, and behavior (e.g., Kahneman 2011) and that these shortcomings should be seriously taken into account to develop more realistic theories and effective strategies for individual and social change (Thaler & Sunstein 2007; Persson & Savulescu 2012).

According to a bold version of this view, recent empirical evidence suggests that we should not push too hard in trying to change or mitigate the influence of distorting factors on human

[40] I will sometimes refer to the idea of moral progress (rather than to mere moral change) only 'indirectly', e.g., to indicate aspects of the descriptive theory of moral change underlying specific accounts of moral progress.
[41] Valuable exceptions to this trend can be found in Buchanan (2020), Buchanan & Powell (2015, 2018). For a more pessimistic view on the possibility of psychological moral change in light of recent empirical research, see Klenk & Sauer (2021).

reasoning and behavior because such an effort would be basically pointless: some aspects of human psychology are just too rigidly insensitive to change (see e.g., Schwitzgebel & Cushman 2015; Haidt 2001; 2012; Kahneman 2003; 2011; Klenk & Sauer 2021; Persson & Savulescu 2012; 2017; Sauer 2019; 2023). Cognitive processes and biases have often been depicted as persistent and hardly educable, like visual illusions: even if we try – in light of our knowledge of how things actually are – we cannot change our perceptual responses (Kahneman 2003)[42]. According to a more moderate interpretation of this view, although active efforts of bias reduction might be slightly effective, they would still be insufficient to cope with some of the central moral challenges of our time (Persson & Savulescu 2017).

Both formulations of this view raise essential challenges for any theory of moral progress that aims to be at least moderately naturalistic, i.e., consistent with empirical evidence and scientific knowledge. As stressed in chapter 1 (especially sec. 1.2, 'A meta-theoretical framework'), a theory of moral progress should rely on a sufficiently realistic descriptive theory of moral change: on the one hand, postulating unrealistic psychological capacities for human beings might favor bad explanatory 'retrospective' hypotheses for social change, that should be rather understood as resulting from more complex or different causes and dynamics. On the other hand, more 'prospectively', a principle of minimal psychological realism would require not to demand and expect people to decide and behave according to normative standards that are impossible for them to comply with (Flanagan 1991; Songhorian 2019).

In light of this, several scholars recently suggested that projects for moral reform based on the improvement of reasoning and deliberative abilities are basically doomed to failure (Klenk & Sauer 2021). Rather than trying to improve people's judgmental and decisional capacities by reducing biases or self-interested motivations, these critics maintain we should bypass the former by leveraging on the latter in an intelligent way that produces the most desirable outcomes (Banaji & Greenwald 2013; Sauer 2019; Thaler & Sunstein 2007). Based on similar premises, a provocative strand of empirically-informed applied philosophy recently stressed that since ordinary means for moral education are not enough to foster significant moral change, we should favor direct pharmacological, neural, or genetic interventions to enhance the moral capacities of individuals (Douglas 2008; Persson & Savulescu 2008; 2012), bypassing explicit moral deliberation (cf. Reichlin 2019).

[42] While biases have been often described by scientists being like visual illusions, it does not mean that intuitive biases are in every way like visual illusions. Having a decision bias may be like being susceptible to a visual illusion in that, in both cases, one can't just think one's way out of it. But decision biases, at least some of them, are products of learning and can be at least partially unlearned.

Part II shows that these projects rely on several false empirical assumptions. After an introductory methodological note on moral change (4.1), in sections (4.2-4.3) I introduce a central empirical assumption at the core of many of these 'ameliorative' projects: the idea that human moral and prosocial cognition and motivation are rigidly doomed to be limited and biased because of the evolutionary history of our species. In section (4.4), I present two recent empirically-informed accounts and strategies for moral improvement that emblematically rely on this assumption. In the following chapters (5-6), I provide evidence and arguments against the hard-wiring thesis. I report critical considerations about the idea that a specific capacity for moral cognition exists and that it should be considered an adaptation – i.e., that was selected because of, and still would perform, a clear fitness-enhancing evolutionary function. I show that recent research and available evidence from several disciplines suggest that human cognitive and motivational infrastructure is not naturally and universally fixed (and biased) as many have recently claimed, and that several socio-cultural means for significant psychological change are possible.

In the last chapters of Part II (7-8) I suggest that important aspects of human morality and moral cognition – above all, the ability for open-ended moral reasoning and normativity – should be conceived as cultural byproducts of the selection of several (both domain-general and domain-specific) psychological traits, which can allow for their development and increased exercise.

4. Hard-wired psychology and moral change

1. How do morals change?

As stated above, the study of moral change will be approached in this chapter with a descriptive temperament, suspending evaluative judgments about the goodness or (un)desirability of more or less specific moral shifts. From this perspective, moral change can be understood as a historical process consisting of significant modifications in social structures and institutions, as well as individual and collective shifts in moral beliefs, attitudes, and behavior (Bloom 2010). On the one hand, the former (supra-individual) kind of moral change consists of shifts in norms, practices, and policies in formal and informal institutions, as well as in the distribution of people across social roles and space (Anderson 2010; Enos 2017; Henrich 2020).[43] On the other hand, moral change also occurs at the level of individual psychology, both within a lifespan and between generations of individuals. This second kind of change includes early moral development, as well as later shifts in moral and non-moral beliefs, attitudes, motivations, abilities for cognitive and emotional control, reasoning, decision-making, and justification (see Part III; Buchanan & Powell 2018, 54–58; Campbell &

[43] I am grateful to Victor Kumar for helping me conceptualize structural moral change in these terms. It is worth noting that several scholars conceive of this kind of supra-individual moral change as the expression or embodiment of collective morality (Macklin 1977, 376; Musschenga & Meynen 2017, 5). However, I see this view as problematic since there is often a significant mismatch between people's beliefs, attitudes and behavior and the values embodied or realized by formal and informal institutions, even in democratic contexts. In my opinion, this is a strong reason for operating and emphasizing the distinction between these two dimensions of moral change.

Kumar 2012; Songhorian et al. 2022). It is important to emphasize that this second kind of change – *psychological* moral change – does not occur within individual minds in isolation, but it is always the product of complex inter-individual, group and cultural dynamics (see Campbell & Kumar 2012; Mercier & Sperber 2017; Sauer 2017).[44]

Table 1. *Dimensions of moral change*

Structural – Institutional	**Psychological – (Inter-)Individual**
• Norms and policies in formal institutions • Norms and practices in informal institutions • Distribution of people in space, social roles	• Beliefs • Attitudes • Motivation • Reasoning • Decision-making

Of course, these two 'kinds' or 'levels' of moral change interact deeply. However, despite interdependencies and significant overlapping between the two, psychological and structural-institutional changes do not always match nor co-occur, and several tensions can persist between these dimensions. For instance, structural power relations often continue to operate even when the values of large numbers of people change (Haslanger 2015). People's attitudes, beliefs, and motivations do not always reflect or align with the values embodied by social and political institutions; also, change in these latter is seldom directly realized by laypeople nor does it perfectly reflect shifts in values, attitudes, or beliefs of the many. For these reasons, considering psychological and supra-individual moral change as distinct objects of analysis regulated by different – though interdependent – dynamics can be useful to study the causal relations between the two. This can help to identify how to promote desired social and moral reforms by acting either on one or the other level in a more informed and effective way.

In discussing core issues within the debate between methodological structuralists and individualists in social change theory, Madva (2016) correctly criticizes oversimplified views according to which either structural or psychological change should *always* have a priority in

[44] Discussion of these two kinds of moral change can be found in the recent philosophical literature on moral progress (see e.g., Buchanan & Powell 2018; Sauer et al. 2021; Songhorian et al. 2022) but, as discussed in Part I, it can hold for more neutral, descriptive analyses of these phenomena as well. In this respect, for instance, Musschenga & Meynen (2017) identify i) *cognitive*, ii) *affective*, and iii) *behavioral* components of psychological moral change. I think this tripartition is misleading. Above all, as far as i) and ii) are concerned, recent empirical and theoretical advancements in moral psychology emphasized the limitations of sharply distinguishing between affective and cognitive processes in moral cognition (see e.g. Bina 2022; Cushman 2013; Saunders 2016).

the promotion of social reform. According to Madva, which of these levels (if any) should be prioritized – either for understanding/explaining or promoting social change – often depends on the specific problems at stake and on the results that one aims to achieve. In line with this view, the aim of these chapters is not to defend the priority of psychological over structural moral change. Its aim is rather to show, against the skeptics, that psychological moral change is possible – even, of course, as a result of structural and other supra-individual changes.

2. The hard-wiring thesis

Within recent empirically-informed debates in practical philosophy, as well as in the social and behavioral sciences, several scholars have claimed that the psychological changes that would be required to face specific collective problems or to comply with certain normative standards are virtually impossible for human beings, at least via ordinary socio-cultural means for social and moral reform such as moral education and reasoning. To recall an already mentioned analogy, according to these views, cognitive biases are persistent and insensitive to change like visual illusions[45]: as we cannot see them differently, often humans cannot perceive, feel, think or behave in certain ways *even if they try* (Klenk & Sauer 2021; Persson & Savulescu 2017; for a critique, see Campbell & Kumar 2012).

One of the main reasons advocates of this pessimistic claim offer in its favor is that the human brain and cognitive infrastructure are significantly shaped and constrained by the evolutionary history of our species (Barkow et al. 1992; Haidt 2012; Haselton & Nettle 2006; Persson & Savulescu 2012; 2017; Sauer 2019; 2023; Street 2006; see also Buchanan 2020 for further critical discussion)[46]. Specifically, psychological changes in beliefs, attitudes, and motivation that would be required to effectively address some of the most urgent moral and cooperative problems of our time – e.g., intergroup conflicts, global poverty and inequality, climate change, and other kinds of existential risk – are virtually impossible because human cognition is doomed to be biased, exclusive, tribalistic and myopic. Ordinary means of moral reform can only be slightly effective, require too much time to produce desired effects, and have no guarantee of success (Klenk & Sauer 2021; Persson & Savulescu 2012).

[45] A commonly cited example is the Müller-Lyer illusion, consisting of two parallel lines of the same length normally perceived as different; after realizing that the lines are equally long (even by measuring them directly), people continue to perceive them as different. Cognitive scientists like Fodor (1983) and Kahneman (2003) see this as evidence that often explicit beliefs and reasoning cannot penetrate our perceptive and intuitive processing.

[46] Boudry et al. (2020) criticize a similar view referred to the limits of potential improvement in theoretical and scientific knowledge and understanding. FitzPatrick (2015) argues that there is no reason or evidence to maintain that human moral capacities and understanding would be constrained by evolution more than other kinds of non-moral knowledge are.

Buchanan and Powell (2015; 2018) recently distinguished advocates of this view into 'Evo-conservatives' and 'Evo-liberals'. Evo-conservatives believe that human nature is essentially impossible to change and appeal to evolutionary explanations to defend pessimistic and conservative views about social and moral change (Arnhart 2005; Asma 2012; Fukuyama 2002; Goldsmith & Posner 2005; Haidt 2012). In particular, they claim that human psychology cannot be stretched over a certain limit and that both normative theories and socio-political action should be adjusted to this fact in order to be feasible (or realistic):[47]

> the ecological challenges our distant ancestors faced generated selection pressures for evaluative tendencies that circumscribe effective moral commitments to members of one's own kin, group, tribe, or nation – and that these putative facts about human evolutionary history significantly constrain the shape of plausible moralities and the scope of other-regarding concern. This, in turn, is thought to suggest that cosmopolitan and other inclusivist moral principles are not appropriate or realistic for beings like us (Buchanan & Powell 2015, 44).

Based on the same empirical premise, other scholars have drawn opposite normative conclusions. Unlike Evo-conservatives, Evo-liberals believe that human moral psychology and behavior can be – *even radically* – modified. This leads them to conclude that psychological moral change should be promoted when needed, though any conceivable 'ordinary' socio-cultural means for moral and social reform currently available is a suboptimal way to do that. To effectively face the main moral mega-challenges of our time, we should rather invest in alternative, empirically-informed strategies for psychological moral change. The most emblematic example of this second perspective is the recent defense of the need for *moral bio-enhancement* (Crutchfield 2021; Douglas 2008; Persson & Savulescu 2008; 2012; 2017). One of the core premises behind this project is that, at the current state of affairs, human psychology is *unfit for the future*: the world has radically changed from the socio-ecological conditions of the Pleistocene, in which our cognitive infrastructure evolved and assumed its current configuration. But new scenarios raise novel moral challenges and normative demands. And, to date, "human beings are not by nature equipped with a moral psychology that empowers them to cope with the moral problems that these new conditions of life create" (Persson & Savulescu 2012, 1).

[47] For more on this point, see Flanagan (1991) and Songhorian (2019). For a critique of the thesis that normative demands should be adjusted to our evolved psychological constraints, see Buchanan & Powell (2015, 65–67).

Following Persson & Savulescu (2017), I call the basic empirical assumption shared by both Evo-conservatives and Evo-liberals the 'hard-wiring' thesis.[48] The main idea behind the hard-wiring thesis is that human psychology is doomed to shortsightedness and tribalism because limited cognitive and prosocial traits were functional to enhance cooperation and benefits within small groups of individuals in the environment of evolutionary adaptedness (EEA)[49], while farsighted and inclusive ones were not. In the next section, I discuss in more detail the evolutionary rationale behind this thesis.

3. Evolutionary explanations of (limited) moral and prosocial cognition

The belief that adaptations to remote human environments and challenges still currently largely influence cognition and behavior is a core thesis in evolutionary psychology (hereafter, EP. See Barkow et al. 1992; Buss 2019; Carruthers et al. 2005; McDonalds et al. 2012; Pinker 2002; Samuels 1998; Tooby & Cosmides 2005; Tooby 2020) as well as a common claim in recent naturalistic approaches to theoretical moral psychology and ethics (Haidt 2012; Joyce 2006; Kitcher 2011; Persson & Savulescu 2012; Sauer 2019; 2023; Singer 2005; Severini 2021; Street 2006).

According to EP, animal cognition and behavior should be understood as products of adaptations to past environmental cues, as are other biological traits such as organs, bones, the immune system, or the capacity of spiders to weave nets.[50] More specifically, EP extends to the explanation of animal cognition and behavior a selectionist-adaptationist approach typically relying on an *etiological theory of function* (Buller 1998; Luco 2019; Kitcher 1993; 2011; Millikan 1984; 1989; Neander 1991). Following Larry Wright (1976), the *function* of *x* is the effect that *x* has in a system, and that accounts for why *x* subsists in it. According to a more moderate conception, however, we can simply conceive of etiology as sufficiently powerful evidence to explain certain dispositional properties. Hence, according to EP, cognitive and

[48] Buchanan (2020) also makes wide use of this expression to challenge a similar idea. He comments: "We often hear that the Darwinian revolution dethroned the dogma that human nature is fixed, unchanging. Yet a version of the dogma still stubbornly persist, ironically clothed in evolutionary garb, even among some of the best evolutionary scientists […] I'm referring to the belief that human nature is fixed in this sense: human morality is essentially tribalistic, group exclusive, and that's not going to change […] The dogma is also sometimes formulated as follows: morally speaking, we are hardwired (or programmed) for tribalism" (1-2).

[49] The EEA is the ancestral set of environmental conditions and selective pressures in which our species is supposed to have adapted; specific to the current discourse, in which the basic elements of human cognition and morality were supposedly selected – somewhere between 1.8 million and 10,000 years ago (Barkow et al. 1992; Buss 2019).

[50] The research program of evolutionary psychology is one of the closest descendants of sociobiology: E. O. Wilson famously claimed that not only animal behavior, but also human morality and ethics should be "biologicized", i.e. explained, understood and improved thanks to the advancements in the life sciences (Wilson 1975, 562).

behavioral traits perform a function when they keep producing effects that causally contribute to their persistence. This very popular theory justifies a common inference in the contemporary debate in moral psychology and philosophy of biology, that draws conclusions about moral cognition and morality's current features and limitations by referring to the evolutionary dynamics and circumstances that explain why they were selected in the past.

Etiological theories of function can be applied to the explanation of several entities, like biological traits, social institutions, artifacts, and more. For instance, noses have the function of smelling because smelling is an effect of noses that causally contributed to their transmission through generations of organisms; money has the function of mediating exchange because making exchanges easier contributed to the persistence of money, and so forth. Likewise, according to several scholars, moral cognition and behavior evolved (i.e. were selected and transmitted) because of their function of enhancing cooperation within small groups of individuals (Alexander 1987/2017; Casebeer 2003; Curry 2016; Greene 2013; Joyce 2006; Kitcher 2005; 2011; Tomasello & Vaish 2013). Modest prosocial traits, such as dispositions to empathize and cooperate with limited groups of people benefited individuals and groups who possessed them, and this contributed to their transmission and persistence over generations. At the same time, however, ethnographic, anthropological and psychological research report that human cognition is not as good at sustaining cooperation *between* groups (Böhm et al. 2020; Greene 2013; Tomasello 2016).

By combining available evidence with this etiological theory of function, evolutionary psychologists hypothesize that several biases, social heuristics and psychological traits such as in-group favoritism, conformism, scope neglect, limited moral concern (etc.) that humans still widely display nowadays have been selected throughout evolution because they contributed to enhancing the reproductive fitness of our ancestors (Haselton & Nettle 2006). In the environment of evolutionary adaptedness (EEA), limited prosocial traits likely had the advantage of protecting early humans from several existential threats, e.g. to physical integrity – from infectious diseases or personal violence – and material resources (Choi & Bowles 2007; Faulkner 2004; Kurzban & Leary 2001; Navarrete & Fessler 2006; McDonald et al. 2012; Neuberg & Schaller 2016; Oaten et al. 2011; Persson & Savulescu 2012, 38; Tomasello 2016).

According to EP (and to accounts relying on its basic claims), the reason why greater prosocial, inclusivist and cooperative dispositions and more sophisticated, unbiased and farsighted deliberative skills and institutions have not been selected for is that their existential costs would have significantly exceeded their benefits in the EEA. Human beings possessing too inclusive, reflective or longtermist attitudes were likely eliminated in the Pleistocene, while

more parochial, intolerant, (selectively) hostile, decisive and impulsive individuals survived; cognitive mechanisms and behavioral responses that allowed humans to deal more efficiently with life-or-death environmental and social challenges were transmitted over generations and selected, while slow and inefficient ones were not, because their bearers had lower chances of survival (Haselton & Nettle 2006).

In other words, according to EP, moral cognition and morality essentially evolved as intra-group, short-sighted and short-termist social technologies. This view, of course, is neither recent nor isolated. Authoritative scholars from different scientific domains have defended similar positions in the past decades. For instance, in *The Limits of Altruism*, ecologist Garrett Hardin stated that morality cannot but exist "on a small scale, over the short term, in certain circumstances and within small, intimate groups" (1977, 26). In *The Selfish Gene*, ethologist Richard Dawkins claimed that "Much as we might wish to believe otherwise, universal love and the welfare of the species as a whole are concepts which simply do not make evolutionary sense" (1976, 2–3). Combine these views with the idea that functional explanations are usually associated with traits' rigidity and resistance to change (cf. Buchanan & Powell 2015, 47), and we have a case for a hard-wired parochial, exclusivist and shortsighted social and moral psychology. On these grounds, evolutionary psychologists have explicitly expressed skepticism about the possibility of significant mental plasticity (Barkow et al. 1992, 39; Pinker 2002). On the same basis, several thinkers have also recently questioned the possibility of robust psychological moral change (Asma 2012; Haidt 2012; Sauer 2019; 2023; Persson & Savulescu 2017; for a review, see Buchanan & Powell 2015).[51]

Note that the fact that some traits were selected because of the advantages they conferred to generations of our ancestors in the EEA by no means implies that these traits are still adaptive now (evolutionary psychologists tend to agree on this point). On the contrary, traits evolved in radically different circumstances can be even very counterproductive and dangerous in modern environments and societies. This problem is technically known as *evolutionary mismatch*: "Our minds are adapted to the small foraging bands in which our family spent ninety-nine percent of its existence, not to the topsy-turvy contingencies we have created since the agricultural and

[51] One of the main claims of classic evolutionary psychology is that the mind is modular: "an array of psychological mechanisms (modules) that is universal among Homo sapiens" (Symons 1992, 139), selected for their adaptedness in the EEA. To use a famous metaphor, according to this view the mind is like a Swiss Army knife: a combination of different tools performing specific functions. For evolutionary psychologists, modules of the mind are many, innate and domain-specific. They are also 'informationally encapsulated', i.e. impenetrable by the cognitive processes generally involved in the activation of different modules of the mind. The modularity of mind hypothesis was first formulated by Fodor (1983), though he conceived only 'low-level' peripheral cognitive systems as modular (e.g. vision and perception). Fodor's hypothesis has been later extended to other cognitive systems – the *massive modularity* hypothesis – by Tooby & Cosmides (1992; 2005), Pinker (1997), Samuels (1998). For critiques of this hypothesis, see Prinz (2006), Sterelny (2010), Pietraszewski & Wertz (2022).

industrial revolutions" (Pinker 1997, 207; see also Li et al. 2017)[52]. To say it with a famous motto, our modern skulls house a stone-age mind (Cosmides & Tooby 2005).

Given these considerations, skeptical views about the feasibility and effectiveness of traditional strategies for moral reform have been offered to the philosophical debate. To illustrate some of the possible practical implications of this skeptical temperament, in the next section I discuss two recent accounts which emblematically rely on the hard-wiring thesis. As stated above, I will not challenge either the ethical-philosophical content nor the moral implications of these accounts. My aim here is rather to show that the theoretical and practical conclusions these accounts draw are untenable, since one of their main empirical premises – the hard-wiring thesis – is empirically unsupported.[53]

Before proceeding, a final consideration about the meaning and scope of evolutionary explanations of morality is in order. Recall that advocates of the hard-wiring thesis claim that biases and tribalistic attitudes are so hard-wired in our cognitive architecture that it is virtually impossible to overcome or mitigate their influence, at least with ordinary socio-cultural tools such as education or reasoning. These evolved shortcomings, they argue, constrain not only (i) human psychology and motivation, but also (ii) moralities – conceived as historically realized systems of practices – and (iii) the content of moral beliefs as well as that of more systematic philosophical ethical theories (see e.g. Haidt 2012; Greene 2007; Schwitzgebel & Cushman 2015; Persson & Savulescu 2012; 2017; Street 2006). According to evolutionary psychology, in fact, even culture and reasoning are the direct "product of evolved psychological mechanisms" (Tooby & Cosmides 1992, 24; see also Haidt 2012; Mercier & Sperber 2017; Sperber 1996; Street 2006).

Advocates of moral hard-wiring, like Persson and Savulescu acknowledge that human morality cannot be fully explained by evolutionary accounts. In fact, they say, morality "is also formed by socio-cultural factors and reasoning [...] But while evolutionary considerations cannot explain these cognitive or doxastic elements of morality [...] they can explain a number of motivational or non-cognitive elements of our moral psychology, and why they can have a hard time catching up with the former elements" (2017, 290). However, as recent experimental

[52] However, some scholars emphasize that several cognitive traits and mechanisms that many see as the problematic legacy of human adaptation to radically different environments are still very adaptive nowadays (Gigerenzer 2007; Gigerenzer et al. 2011; Page 2021; Railton 2014; 2017).

[53] Another, more moderate way to frame this problem may be to state not that human moral psychology is either hard-wired or it is not, but rather that the influence of our evolved traits and biases (and of our evolutionary history more in general) on human cognition and behavior is more or less significant (so disagreement would be about grade rather than being an 'all or nothing' issue). Of course, however, it would be hard to quantify such an influence (10, 50, 80%?). Hence, for argument's sake, I will stick to the hard-wiring formula; but the unconvinced reader could substitute 'hard-wiring' with 'steep-climbing'. Thanks to Josh Greene for making me clarify this point.

research points out, cognitive and doxastic elements of morality often depend on 'non-cognitive' or pre-reflective cognitive processes (Bago & De Neys 2019; Bialek & De Neys 2017; Campbell & Kumar 2012; Cushman 2013; Damasio 1994; Greene et al. 2001; Greene 2007; Haidt 2001; 2012; Prinz 2006; 2007).[54]

Hence, if we accept Persson and Savulescu's claim that evolutionary explanations mostly account for motivational and 'non-cognitive' aspects of our moral psychology (i), then we should also accept that 'cognitive' or doxastic aspects of morality (ii-iii) might, at least partially, be explained by evolutionary considerations. Unsurprisingly, in several works the authors state that evolved natural inclinations – such as the universal human preference for indirect harm – significantly affect the very content of our ethical theories (Persson & Savulescu 2012; 2017; see also Cushman 2013; Cushman et al. 2006; 2012a; 2012b; Rozyman & Baron 2002; Singer 2005). If this were so, then the fact that moral cognition is still shaped, nowadays, by its being selected to face socio-environmental cues in the EEA would rigidly constrain not only motivational aspects of morality, but also the content of moral beliefs, principles, and theories (see e.g. Greene 2007; Street 2006). If this view were correct, the possibility of radical moral change (at all the aforementioned levels) would actually be considerably reduced.

4. Moral change *despite* hard-wiring

As mentioned in Section 3, the hard-wiring thesis does not necessarily imply conservative conclusions for individual and societal moral change. On the contrary, Evo-liberals have argued that the hard-wiring thesis and the possibility of significant moral change are, in actuality, compatible. Traditional efforts to change people's minds and behavior towards greater inclusivity, prosociality, increased moral motivation or moral reasoning abilities – such as education and argumentation – turned out to be generally ineffective. However, alternative strategies for moral reform can be way more impactful, and should be seriously taken into consideration, especially in light of new scientific and technological advances.

One of the most challenging and debated among these alternative strategies is the already cited project of moral bio-enhancement. According to some of its main proponents, traditional methods of moral education and reasoning "have had modest success during the last couple of millennia" (Persson & Savulescu 2012, 9), and since evolved "limitations of human moral

[54] On the use of the word 'cognitive', see Greene (2007, 40-41).

psychology pose significant obstacles to coping with the current moral mega-problems [...] biomedical modification of human moral psychology may be necessary" (Persson & Savulescu 2017, 287).[55]

Based on similar assumptions, a more moderate 'compatibilist' view that tries to hold together both the hard-wiring thesis and the possibility of (progressive) moral change has been proposed by Hanno Sauer (2019). According to Sauer, even if human cognitive capacities are significantly constrained by evolution, "we should not expect our individual moral psychology to play an important role in promoting or maintaining moral progress" (158);[56] on the contrary, we should "bypass, rather than further stretch, the constraints of our evolved psychology to make moral progress possible" (153), working around it with "clever institutional kludges" (163).

Sauer calls this mechanism 'institutional bypassing' (162). It relies on the idea that important moral shifts occur without involving significant change in peoples' minds – e.g. in their abilities for information-processing, imagination, conceptual understanding (cf. Moody-Adams 1999; Severini 2021), consistency reasoning (Campbell & Kumar 2012) or justification (Songhorian et al. 2022), empathic imagination and regulation, bias reduction (Schaefer & Savulescu 2019), and so forth.

According to this view, the extraordinary moral change that societies have been experiencing over (especially recent) history mostly depends on, and consists in the development of supra-individual institutions able to better secure social cooperation, rather than involving changes in people's minds:

> Modern humans are hypersocial in a way that cannot be attributed to changes in our psychological capacities because these developments are too recent. [...] There are evolutionary limits to human inclusiveness. Does this mean that evolutionary conservatives are right, and large-scale cooperation is not a feasible political ideal? Of course not. There is an institutional arrangement – the market – that provides a workaround. It facilitates extensive chains of cooperation, and incentives to benefit others, without relying on the baker's (or anyone's!) benevolence (Sauer 2019, 163; see also Sauer 2023, Introduction and chapter 4).

As stated, I am not interested in discussing the *ethical* premises and implications of these accounts of moral progress here. I am mostly concerned with whether the empirical thesis these accounts rely on is correct. In the following sections, I review evidence and offer arguments

[55] See Persson & Savulescu (2012), 118-121.

[56] "Consider the often vast transformations many societies have undergone over the past decades or centuries. How many of these, and to what degree, can plausibly be attributed to changes in individual people's psychology which likely remained unaltered over the course of such relatively recent and swift developments?" (Sauer 2023, 3.5)

suggesting that the hard-wiring thesis – and the theories of moral change that rely on it, such as the two just briefly considered – contrast with empirical evidence.

A potential response to my objections might be that they are based on a misunderstanding of the hard-wiring thesis. According to this reply, Evo-conservatives and Evo-liberals do not claim that human psychology is permanently fixed and completely insensitive to change. Evo-liberals especially do not deny that psychological change is in principle possible. More modestly, what they stress is that significant psychological change would require too much time and effort to be sufficiently widespread, stable, and effective for addressing some of the main contemporary moral mega-problems. The evidence I review in the following sections is ordered from the longest to the shortest time frame required for psychological moral change to occur. This evidence suggests that several socio-cultural drivers can significantly shape human moral cognition and behavior both over centuries and within one single life.

5. Against moral hard-wiring

1. Kinds of evolutionary explanations

A first objection to the hard-wiring thesis stems from a general skepticism towards the explanatory power of methods and hypotheses of classic EP. Specifically, some perplexities are addressed to the reliability of EP's theoretical tools in understanding social and moral cognition in ecological, material, structural, and epistemic conditions which radically differ from the EEA (i.e. the Pleistocene). A comprehensive critique of EP lies, however, beyond the scope of this work.[57] I will thus limit my considerations to the idea that human moral and prosocial cognition and behavior could be understood by referring to core methodologies, assumptions, and theoretical claims of classic EP.

How can we assess whether and how some definitory traits of a *specifically moral* kind of cognition technically evolved? As stated above, EP approaches the study and explanation of cognitive and behavioral traits as other branches of evolutionary science (e.g. evolutionary biology) seek to explain the evolution of biological traits. Biological traits can evolve in several ways: for the current purposes, let us focus on two of them: *adaptations* and *exaptations*.

As discussed in the previous section, evolutionary psychologists tend to conceive of morality and moral cognition as evolutionary adaptations, i.e. as being directly selected by

[57] One of the most extensive critiques of evolutionary psychology has been formulated by Buller (2005). See also Gould & Lewontin (1979) for a classical critique of the abuses of adaptationist/selectionist explanations of biological, psychological and behavioral traits. According to Gould and Lewontin, evolutionary psychologists often make up unproven 'just-so-stories' tracing back current traits to functional responses to environmental challenges that our ancestors would have faced in the Pleistocene.

evolutionary pressures because of their contribution to enhance cooperation and reciprocal benefits among individuals and groups, hence increasing chances of survival and enhancing reproductive fitness. Adaptive traits get selected when i) their (random) appearance and transmission by genetic heritability turns out to be fitness-enhancing, i.e. its bearers live longer than those who lack it, and leave more descendants to which the trait is transmitted, and ii) their main function keeps being the function that contributed to their natural selection.

On the other hand, *exaptations* are evolutionary by-products. Some traits that get selected because of their fitness-enhancing contribution – e.g. because they facilitate addressing more efficiently specific socio-ecological challenges – can later turn out to be useful to face even radically different problems, for which they are subsequently recruited. In one of the most famous papers in the history of evolutionary theory, Gould and Lewontin (1979) referred to this phenomenon as "secondary adaptation" (596), or "the fruitful use of available parts" (584). The term *exaptations* was later coined by Gould and Vrba (1982), who defined them as traits "evolved for other usages (or for no function at all), and later 'coopted' for their current role" (6).[58]

Hence, only adaptations can be understood, strictly speaking, as the 'direct' result of natural selection. How do we know whether a trait is an evolutionary adaptation, i.e., if it was selected because of its fitness-enhancing contribution? Some relatively uncontroversial indicators can help to find out whether or not a trait can be considered an adaptation. Let's briefly examine a few of the most important ones. First, there are good reasons to hold a trait an adaptation if we can clearly identify a *specific function* that explains its selection, transmission and persistence over generations of organisms. As mentioned above, for instance, as far as cognitive and behavioral traits are concerned, classic evolutionary psychology tends to see the mind as composed of domain-specific modules, each of which can be given clear functional explanations, which can be traced back to adaptive responses to vital environmental challenges in the EEA (e.g. Barkow et al. 1992).

Other important indicators to assess the evolutionary adaptedness of a trait are its antiquity, universality, and early ontogenetic development. These features often co-occur. For instance, traits that are universally shared by members of a species – including psychological ones, such as the ability to learn any possible language (Chomsky 1975) – generally also develop very

[58] For instance, feathers likely evolved in the first place to keep the body warm, and only later were co-opted for flying. Some plants likely used to secrete resins in the first place to defend against herbivores, and only later resins became a reward for pollinators (Garson 2008). Several cultural human activities (e.g. literature) can also be considered exaptations, since they can be conceived of as byproducts of the evolution of other forms of intelligence (see Ayala 2010; Buss et al. 1998; Gould & Lewontin 1979).

early in life. This phenomenon can be explained by hypothesizing either that a) these traits are easily acquired as by-products of the ontogenetic development of other traits (e.g. they are learned after having developed other, domain-general cognitive capacities), or b) that they are innate,[59] i.e. the result of selective pressures operating on specific ontogenetic developmental pathways. According to EP, evolved modules of the human mind satisfy all these conditions: they are evolutionarily old, universal across societies and cultures, and develop very early in life, independently of the acquisition of other psychological capacities (Barkow et al. 1992; Tooby & Cosmides 1990). Hence, they can be considered adaptations.

Now, to evaluate if the hard-wiring thesis is correct – i.e. whether the main function of moral cognition and morality is still that of solving small-group cooperative problems because that is the function for which they have been selected – we may try to assess whether they satisfy the aforementioned conditions for being considered evolutionary adaptations. Is there something such as a peculiar type or set of mental processes or content – if not even specific neural substrates – that are uniquely involved in moral experiences and computation?[60] Is moral cognition universal, ancient, innate, functionally and domain-specific?

2. Are moral cognition and morality adaptations?

Answering this question is a complex task, and significant disagreement among scholars persists on this issue. Further complexity is added, of course, by the problem of unequivocally defining conceptually what morality and moral cognition are – even before trying to identify, empirically, how they manifest and where they might be located in the brain (Young & Dungan 2012). It is, in fact, not clear at all that it is possible to identify unified, consistent, and universal phenomena as morality or moral cognition as objects of inquiry (see e.g. Goodwin 2017; Hindriks & Sauer 2020; Parkinson et al. 2011; Sinnott-Armstrong & Wheatley 2014). Trying to shed light on these issues is of great relevance for the current purposes, since if we cannot legitimately affirm that a specifically moral kind of cognition exists, or that morality can be given a clear functionalist-adaptationist explanation and definition, the idea that human moral cognition and morality are naturally limited by their evolutionary history becomes untenable.

[59] The link between adaptations and innateness is partly problematic since, e.g., innate traits like genetic diseases are not adaptations, while the development of certain adaptive traits requires learning (see Griffiths et al. 2009; Mallon & Weinberg 2006; Samuels 2002).

[60] See Arvan (2021, 96-99), Greene & Young (2020), Han (2017), Pascual et al. (2013), Young & Dungan (2012) for some reviews of experimental findings about the neural correlates of moral cognition.

A relatively non-controversial starting point to understand what moral cognition might be – in order to assess whether and how it evolved – is to treat it as a peculiar kind of *normative* cognition. Empirical evidence, evolutionary models and theoretical research seem to converge on the idea that normative cognition can be legitimately considered an adaptation. Why?

First, the function of 'general' normative cognition seems more straightforward and less controversial. Normative cognition consists in the regulation of one's and others' conduct to maintain norm-conformity, by proscribing and prescribing types of actions, and regulating sanctions and punishment for free-riders (Birch 2021; Kumar & Campbell 2022, chapter 3). So conceived, norm-compliance and the disposition to punish defectors are psychological traits which evolved because they favored and stabilized cooperation and cultural transmission (Birch 2021; Boyd & Richerson 1992; Kumar & Campbell 2022; Machery & Mallon 2010). Birch (2021) identifies three core elements of normative cognition: i) to detect or predict failures in norm-compliance (both by others and oneself); ii) to activate emotional and motivational pressure (e.g. anger, shame, uneasiness) to anticipate or intervene on failures in norm-compliance; iii) the activation of solution strategies to restore conformity (by correcting one's or others' behavior, asking for forgiveness, or punishing).

Second, although it is not easy to ascertain how ancient the appearance of normative cognition and behavior may be, norms and normative cognition seem to be ubiquitous in human societies and emerged very early in their evolutionary history (Kumar & Campbell, 2022; Machery & Mallon 2010). Andrews (2020) suggested that some basic forms of normative cognition and behavior can be found not only in the great apes, but also in other non-human animals (especially mammals) who are able to distinguish a) agential vs. non-agential violations of norms, b) in-group vs. out-group normal behavior (choosing to follow what the majority of one's group does), and c) negatively reacting in front of others' violations of norms and implementing corrective strategies.

Third, paradigmatic experimental studies report that a distinctive, domain-specific capacity to detect cheaters and norm violations respects the remaining aforementioned conditions for considering an evolved capacity an adaptation (Cosmides & Tooby 2005; Cummins 1996b; Machery & Mallon 2010). In particular, these studies show that both adults and very young children perform significantly better in reasoning tasks that involve the violation of *deontic* conditionals such as "if X is here, then Y *must be* there" than in tasks involving *indicative*

conditionals such as "if A is here, then B *is* there" (see Cummins 1996a; 1996b; Harris & Núñez 1996).[61]

Nonetheless, we might doubt that the cognitive processes at play in an alleged *specifically moral* kind of cognition are the same as those at play in any other kind of normative cognition. There are in fact several kinds of norms to which we (reasonably) do not refer as 'moral', and others that are not morally relevant at all.[62] Although several psychological traits involved in moral cognition *as in other domains* – like norm-based cognition and emotions – can be legitimately considered evolutionary adaptations, the idea that a specific capacity for moral cognition was *directly* selected by evolution is not supported by available research and evidence (Arvan 2022; Machery & Mallon 2010). If we consider the aforementioned indicators to hold a capacity an evolutionary adaptation – functional specificity, universality, antiquity, innateness –, none of them seems to be clearly present in the case of moral cognition.

Unclear and contingent function. Although many functionalist-adaptationist accounts of moral cognition and morality have been defended in recent years (Casebeer 2003; Cosmides & Tooby 1992; Curry 2016; de Waal 2006; Dennett 2003; Joyce 2006; Kitcher 2011; Luco 2014; 2019; Rai & Fiske 2011; Sinclair 2012; Tomasello & Vaish 2013), the task of identifying the biological-evolutionary function of morality and of a specific kind of cognition dedicated to morality is particularly problematic (Buchanan & Powell 2015; 2018, 79–91 and 387-388; Smyth 2017).

The most common view among adaptationist accounts is to conceive of the main function of morality and moral cognition as that of reducing social conflicts by fostering cooperation and social cohesion and, according to some, even of increasing well-being and its equal distribution (Boehm 2012; Luco 2019; Railton 1986). Certainly, this view makes sense. Moral institutions (e.g. systems of moral norms) may have likely been useful for dealing with several socio-environmental needs and challenges over the course of evolution, and cognitive traits involved in moral thinking and reasoning might have been selected because they contributed to benefit both individuals and groups who displayed them (Baumard et al. 2013; Boehm 2012; Buchanan 2020, 138-142, 152-153; Campbell & Kumar 2012; Campbell & Woodrow 2003, 361-371; Luco 2019; Stanford 2019, 11, 19-20). However, to reduce moral cognition and morality to the overly simplified functions that adaptationist accounts usually refer to – i.e.

[61] For more on why normative cognition should be considered an evolutionary adaptation, see Machery & Mallon (2010, 11-20), Kumar & Campbell (2022, chapter 3).

[62] Other, non-moral (or only indirectly morally relevant) kinds of normativity may concern: logical consistency and/or rationality (e.g. "given *p*, you should/shouldn't conclude/believe that *q*"), aesthetics ("never wear plaid and stripes together"), prudence ("if you want to be healthy, you should eat better/do sports"), or conventions ("you should always leave a tip here").

enhancing cooperation, reducing social tensions, promoting well-being – seems incorrect for several reasons.[63]

First, if to endorse a functionalist-adaptationist account of moral cognition and morality means to state that the function(s) they perform now is/are the same ones for which they have been selected over the course of evolution, this claim is both epistemically unjustified (Godfrey-Smith 1994; Smyth 2017) and empirically false. As far as the empirical plausibility of these views is concerned, recent historical trends, cross-cultural empirical research, and comparative ethological studies show that different environmental, social, and epistemic conditions predict not only change in moral institutions, norms, and collective practices, but also in individuals' psychological capacities involved in moral reasoning and deliberation, sensitivity and behavior, highlighting the extraordinary plasticity of humans' psychological outlook (Buchanan & Powell 2018; Kumar & Campbell 2022; see also Reger et al. 2018; Stoks et al. 2016; Watkins 2020, for adaptive plasticity in non-human species). Specifically, a massive amount of empirical data show that humans can reason, feel, and act by significantly deviating from what would be prescribed by biological and socio-cultural 'functional' normativity depending on the socio-ecological circumstances (Buchanan & Powell 2015; 2018; Henrich 2020; Welzel 2013; Inglehart 2018).

Therefore, the way we understand the function of morality and moral cognition can depend on different environments and normative standards that can emerge in response to different environmental cues. Actually, there are never 'absolute' or 'intrinsic' functions, i.e. existing independently of the larger system of which the entity at stake is part; functions always depend on systemic-ecological circumstances, and vary when these conditions change (Smyth 2017, 1131-32).[64] As other biological traits, institutions, or artifacts, certain cognitive traits and moral institutions can be considered more functional in some circumstances than in others, and different socio-environmental conditions can radically alter what is more useful, efficient, or appropriate to assess specific environmental problems, even by altering phenotypical characteristics in a very short time.

[63] An 'indirect' reason for why morality and moral cognition might not have (had) the function of enhancing cooperation, social cohesion and prosperity that many scholars attribute to them is that there are many, stronger alternative explanatory hypotheses for the higher levels of cooperation, cohesion and prosperity that several human societies experienced in recent history. Progressively increasing levels of well-being and social cohesion, for instance, strongly correlate with progressive increase in the political and repressive power of States and legal institutions (Runge 1984), as well as with the development of market institutions and the spread of capitalism (more on this below).

[64] These environmental circumstances can be conceived of as functions' enabling conditions. For example, the function of polar bears' fur is that of camouflaging them *because* of the whiteness of the environment they live in; in other words, the whiteness of their environment is the enabling condition for their fur's function. If, somehow, a population of polar bears were moved to a different, non-white context, we should conclude that the original function of the bear's fur has been lost (Smyth 2017, 1132).

Although the function of moral cognition and morality appears to be more complex, contextual, and historically variable than many have claimed (Smyth 2017), some scholars emphasize that several elements of the ancestral conditions in which proto-morality and moral cognition evolved are still present in contemporary societies (Greene 2013; Kitcher 2011). Philip Kitcher, for instance, conceives of "the current human situation as analogous to that initially prompting the ethical project. As it was in the beginning, so too now – for the conflicts to which our ancestors' lives were subject are mirrored in contemporary hostilities across the human population" (Kitcher 2011, 8). According to Kitcher, moral cognition and morality still perform the same evolutionary function for which they were selected. This is, however, only partly true. As I will argue more thoroughly in the following sections, humans have developed social practices, institutions, cognitive abilities, and dispositions that completely disrupted the socio-ecological circumstances in which – according to several prominent scholars – morality and moral cognition were selected because of their adaptedness. In turn, morality and moral cognition's function – assuming that we can identify one – cannot but be changed as well (Smyth 2017).

The fact that ecological conditions change could constitute a problem if our moral cognition were actually rigidly fixed to their original ancestral configurations, and unable to face the challenges that new environmental conditions may require (recall Persson & Savulescu and Pinker's claims about evolutionary mismatch quoted above). As I will stress later, however, this does not appear to be the case. As Joshua Greene and other scholars have recently highlighted, in fact, only certain aspects of our moral psychology (S1, or *model-free* mechanisms) seem to be more directly explainable by reference to their etiology and learning history (not only phylogenetic, but also ontogenetic; not only 'natural', but also cultural) and for this reason they are often more inflexible and harder to control and modify (Cushman 2013; Greene 2017). However, other cognitive mechanisms and capacities involved in moral (as in non-moral) learning and decision-making (S2, or *model-based* mechanisms) are significantly more flexible and amenable to evidence, conscious and critical scrutiny. Relying more on the latter rather than on the former might help us better address new and more complex contemporary moral problems (Bina 2022; Greene 2013; 2014; 2017); and even judgments based on model-free intuitions can change if there are new learning opportunities of the right kind.

Universality. Conceiving moral cognition and morality as universal is also problematic. Of course, this judgment strongly depends on our definition/conception of morality. But although

evidence and theoretical models support the idea that norms and normative cognition are likely present in every human society and culture, it is not equally clear whether *moral* norms and cognition are also universal, even if we assume a very minimal definition of moral cognition as a peculiar kind of normative cognition that is *authority-independent* and involves concern for an *impartial consideration of interests* (see Joyce 2006, 70–71; Railton 2017, 173).[65]

Oldness. Considering the above, it seems implausible also to maintain that moral cognition and morality are evolutionarily ancient, both if we look at the history of humanity and at the cognitive and behavioral traits of our non-human ancestors. Some aspects of human cognition involved in moral judgments and behavior are certainly present in our non-human relatives (Brosnan 2006; Brosnan & de Waal 2006; de Waal 1996). However, without denying any biological continuity nor conceiving human capacities as 'privileged' or 'superior', many comparable cognitive skills and institutions cannot be found in our non-human ancestors (Bowles & Gintis 2013; Henrich 2015; Machery & Mallon 5–10; Prinz 2008, 397–402). Moreover, as we will see more in detail in the next sections, cross-cultural and historical evidence shows that several cognitive traits and configurations of moral institutions appeared in the history of human societies only very recently (Buchanan & Powell 2015; 2018; Henrich 2020; Kumar & Campbell 2022; Pinker 2011; Schulz et al. 2019).

Innateness, domain-specificity, and the moral-conventional distinction. The plausibility of an innate, domain-specific capacity for moral cognition has been widely questioned in recent years. In the past decades, several scholars have defended the idea that humans might have an innate capacity to distinguish moral norms from non-moral ones (e.g. Dwyer et al. 2010; Hauser 2006; Joyce 2006), especially in light of influential studies on the 'moral-conventional distinction' in young children (Smetana 1981; Turiel 1983; for a review, see Machery & Stich 2022). According to this fortunate paradigm, evidence shows that young children cross-culturally distinguish between norms that are seen as independent from authority, universally applicable, and justifiable by reference to harms and rights (*moral* norms), and norms that depend on authority and/or contextual contingencies, and whose violations are judged less seriously (*conventional* norms). According to this distinction, for example, it is *morally* unacceptable to hit people for fun, but it is only *conventionally* unacceptable to go to school or the office wearing flip-flops or pajamas. The ability of children to distinguish between these norms very early in life has been seen by many as strong evidence of the innateness of a peculiar

[65] For more on this point, see Machery & Mallon (2010). As the authors point out, "the universality of norms should not be confused with the universality of moral norms" (31).

kind of normative cognition, i.e., *moral* cognition. This conclusion is based on a kind of 'poverty-of-the-stimulus' argument: such a universal sensitivity to the moral-conventional distinction emerges so early in life that it cannot depend on the development of other cognitive skills. This suggests that it might be unlearned, i.e. innate (Mikhail 2007).[66]

However, several scholars have pointed out that empirical evidence is far from supporting this conclusion (Kelly et al. 2007; Machery & Stich 2022; Machery & Mallon 2010; Sinnott-Armstrong & Wheatley 2014; Stich 2018; Young & Dungan 2012). First, other classic studies on the moral-conventional distinction in young children show that kids almost never conceive of social norms as genuinely 'conventional' (Shweder et al. 1987; Gabennesch 1990). Second, the idea that moral and conventional rules are clearly distinguishable according to the aforementioned criteria of authority-dependence, scope of validity/applicability, harm-sensitivity, and seriousness of violations contrasts with empirical evidence. First, these features often do not co-occur; we can list infinite cases of moral and conventional norms that do not meet one or more of the aforementioned criteria. For example, violations of norms that we might treat as moral can be less serious than conventional ones. Breaking the promise not to eat the last slice of cake at home or telling a white lie seem to be violations of the former case, but we judge them less seriously than, e.g., going to work naked or continuing to interrupt a seminar without raising one's hand. Or consider the norm according to which nobody should eat putrefying meat. This norm appears to be authority-independent, universal, harm-sensitive and serious, but there are no reasons to consider it a moral rule. Furthermore, several studies in social psychology (Cushman et al. 2012; Haidt et al. 1993) report that people judge fake or victimless actions[67] as seriously wrong and authority-independent, but without justifying them by referring to harms, rights, or empathic concern for the subjects involved – since there are actually no victims nor harms involved (for further examples and discussion, see Prinz 2008, 384–385; Machery & Stich 2022; Machery and Mallon 2010, 33–35, Stich 2018).[68]

As Machery and Mallon point out,

> while many philosophers, psychologists, and anthropologists have claimed that morality is a product of the evolution of the human species, the evidence for this claim is weak at best. First, we do not know whether moral norms are present in every culture: because researchers endorse rich characterizations of what moral norms are, it is not obvious that norms that

[66] The poverty of the stimulus argument was originally formulated by Chomsky (1975) to explain the capacity to speak virtually any possible language just being exposed to a few (insufficient) environmental stimuli, before and independently of the development of other cognitive traits.
[67] Such as masturbating with a dead chicken or cleaning the toilet with one's national flag (see Haidt 1993; 2012).
[68] Note that, while there might be borderline cases, these are not conclusive reasons to abandon a conceptual distinction between moral and conventional norms.

> have the distinctive properties of moral norms will be found in every culture, and, in any case, researchers have simply not shown that, in numerous cultures, there are norms that fit some rich characterization of moral norms. Second, the claim that early on children display some complex moral knowledge in spite of variable and impoverished environmental stimuli is based on the research on the moral/conventional distinction. Although this research remains widely accepted in much of psychology, a growing body of evidence has highlighted its shortcomings. Third, the other pieces of evidence often cited in the literature on the evolution of morality do not constitute evidence that moral norms and moral judgments, understood as a specific type of norms and normative judgments, evolved, rather than evidence that normative cognition evolved (2010, 35).[69]

Genetic heritability, culture and cognitive-behavioral development. The idea that individual cognitive development can be conceived as rigidly genetically pre-programmed is also scientifically incorrect. Recent research in behavioral genetics and cultural evolution emphasizes that genetic expression and individual cognitive and behavioral development critically depend on interactions between genes and multiple environmental, social and cultural factors (Henrich 2015; Kumar & Campbell 2022; Stotz 2014; Schulz 2020; Sterelny 2012; Uchiyama et al. 2021). There are indeed good reasons to "reject the predeterministic and instructionist frameworks common to contemporary evolutionary psychology and other gene-centered perspectives on human behavior" (Lickliter & Honeycutt 2003).

Therefore, understanding moral cognition and morality as evolutionary adaptations appears problematic. Can neuroscientific research offer relevant insights on this issue? Is there any neurocognitive mechanism specifically dedicated to moral experiences (a 'moral brain') which can be isolated from, e.g. mechanisms involved in other related kinds of computation and experience – e.g. emotions, empathy, social cognition, norm cognition, cost-benefit analysis, prudential reasoning? Recent neuroscientific evidence shows that the main neural correlates of moral judgment and decision-making recruit several brain systems and pathways involved in many non-moral activities, emotional processing, social cognition, valuation, counterfactual reasoning, agency, cognitive control, planning, mental time-travel, prudential reasoning, and so forth (Arvan 2022; Borg et al. 2006; Cushman 2008; Gray & Wegner 2009; Greene et al. 2004; 2008; Kennett & Matthews 2009; Paxton et al. 2012; Shenhav & Greene 2010; Waytz et al. 2010). None of these capacities, however, is specifically or uniquely involved in moral cognition. Basically, it seems that "moral neuroscience has provided no candidates for

[69] For another skeptical view on moral nativism, see Sterelny (2010).

substrates or systems dedicated to moral cognition [...] So far, the uniquely moral brain has appeared nowhere – perhaps because it does not exist" (Young & Dungan 2012, 5-7).

On the contrary, moral cognition, practices and institutions seem to involve and critically depend on the development and use of several adaptive cognitive traits (Ayala 2010; Nichols 2004; Prinz 2008), but none of them is "distinctly 'moral' or inherently conducive to social cooperation. [...] [T]hey were each plausibly selected in evolutionary history for amoral reasons: as capacities that enable fitness advantages irrespective of whether they are used to general moral actions conducive to social cooperation. [...] [M]oral cognition is almost certainly not a biological adaptation for social cooperation" (Arvan 2022, 99).

In this section, I tried to address the empirical soundness of the hard-wiring thesis – the idea that moral cognition and morality are still bound by the reasons for which they were selected in the EEA – by considering the plausibility of influential functionalist-adaptationist explanations of moral cognition and moral institutions. In particular, I challenged the idea that morality and moral cognition can be conceived as universal and domain-specific sets of cognitive-behavioral traits dedicated to perform specific evolutionary functions. The idea that specifically moral cognitive traits evolved in a way that makes them hard to modify nowadays – because they are still doomed to perform their original, ancestral bio-cultural function – appears problematic. The points just made provide preliminary reasons to cast doubt on the soundness of the hard-wiring thesis.

A second possibility is to consider human moral capacities evolutionary by-products, rather than adaptations. What if morality, rather than being directly promoted by natural selection, was just a by-product of other cognitive and social traits (that were directly selected for their fitness-enhancing contribution)? Although prominent contributors to this debate have categorically rejected this hypothesis (see Buchanan & Powell 2015; Machery & Mallon 2010, 23), I find this view more appealing. I address this problem more directly in chapter 7. Before doing that, let me address some further empirical objections to the hard-wiring thesis.

3. EEA: Population size, intergroup contact and hostility

In the previous section, we saw that the hard-wiring thesis is grounded on assumptions concerning features of human populations and other socio-ecological conditions and challenges in the EEA. Several scholars have inferred from paleoanthropological records and recent socio-psychological research that the cognitive and motivational shortcomings human beings widely display nowadays are the direct product of selective pressures in ancestral environments. These

pressures rewarded and favored the transmission of traits and responses that were advantageous in small societies of hunters and gatherers: "these limitations are the result of the evolutionary function of morality being to maximize the fitness of small cooperative groups competing for resources" (Persson & Savulescu 2017, 286; see also Greene 2013, 23). In this section, I unpack this thesis into three empirical sub-claims and try to assess their validity.

i. Population size and interactions. Human communities have been small and relatively homogeneous (both genetically and culturally) for most of *homo sapiens*' history. Also, available technologies in the EEA allowed limited interactions between individuals and groups (e.g. geographically and culturally distant ones). Chances to harm, benefit, cooperate with, and to morally consider distant and diverse subjects and events have been extremely low for hundreds of thousands of years (Greene 2007; Persson & Savulescu 2012; 2017; Singer 2005). Imagination, anticipation and evaluation of consequences, caring about and investing in cooperative relationships with faraway and/or culturally diverse individuals, deviating from local moral and social norms are all costly behaviors; for thousands of years they have been extremely much more costly than today, and there is no reason why bio-cultural selective pressures should have selected ultra-inclusivist, open-minded, and farsighted psychological traits. If the etiological theory of function introduced above is plausible, this is why humans' cognitive, moral, and prosocial traits are still very biased and limited in scope nowadays.

ii. Tribalism and hostility. In the past decades, research and discussion from several disciplines has emphasized that human cognition is strongly, rigidly, universally biased towards in-group favoritism and different degrees of discrimination against out-groups. According to some studies, such differences in attitudes and beliefs about in-groups and out-groups are often just a matter of 'social discounting' (we just care more about kins and less about strangers). But oftentimes these traits are also seen as closely related to intolerance, disgust, contempt, xenophobia, direct hostility and even aggression (Choi & Bowles 2007; De Dreu et al 2022; Kumar & Campbell 2022; Lee 2016; Wrangham and Peterson 1996).

iii. Selective pressures. Limits to imaginative and decisional capacities, enlarged and impersonal prosociality, and to inclusivist and universalist moralities – several kinds of cognitive biases, conformism, parochial altruism, in-group favoritism, scope neglect, discrimination, hostility (etc.) – have been selected because they have been adaptive for thousands of years in pre-modern societies; i.e., they were fitness-enhancing, either because more efficient or because they contributed to protecting individuals from several existential

threats (Aarøe et al. 2017; Bennis et al 2010; Faulkner et al. 2004; Haselton & Nettle 2006; Kurzban & Leary 2001; Navarrete & Fessler 2006; Neuberg & Schaller 2016; Oaten et al. 2011; Page 2021; Tomasello 2016). On the contrary, more farsighted, open-minded, inclusive, and prosocial traits such as extended intergroup trust, tolerance and cooperation, or openness to new experiences and change were not selected because they turned out to be more costly than beneficial: ultra-cooperative, adventurous, and farsighted traits did not survive, reproduce and spread as much as moderately-cooperative, conformist, and shortsighted cognitive and behavioral traits.[70] In what follows, I show that recent research suggests that these three main claims are empirically unsupported.

Counterevidence

i. First, according to paleoanthropological evidence, it is not clear that early human societies were as small and close-knit as many depict. Archaeological records report that human communities have been composed of thousands of people at least over the past 50,000 years (Richerson & Boyd 1999; Sterelny 2019) – numbers that are far from the small groups of 50 to 150 individuals often mentioned in the literature (Dunbar 1993; 2010; Persson & Savulescu 2012). Even contemporary hunter-gatherer communities – such as the Martu from Western Australian desert, or the !Kung San people from Western Kalahari – stably coexist and cooperate with groups of thousands of people (Richerson & Boyd 1999, 254). Recent evidence about social networks and interactions between residential units in contemporary hunting and gathering communities shows that these societies are organized in large groups of thousands of individuals; and that residency, group membership, and social interactions are much more fluid and dynamic than previously thought, approaching the numbers of modern communities based on agriculture and industry (Bird et al. 2019; Segovia-Cuéllar & Del Savio 2021).

ii. As already suggested by Allport (1954), recent anthropological studies on intergroup relations have questioned the link between in-group favoritism and hostility towards out-groups, showing that no empirical support exists for this correlation (Corr et al. 2015; Yamagishi & Mifune 2016), both among humans and non-human primates (Brewer 1999; Pisor & Surbeck 2019). Moreover, recent research in evolutionary biology and anthropology also

[70] A compatible conclusion is reached by Axelrod (1984) in one of the most classical works that tried to understand evolution and the benefits of different prosocial traits and cooperative strategies for both individuals and groups. Axelrod used game theoretical models simulating potential dynamics of competition between more and less prosocial behavioral traits and strategies. According to Axelrod's study, the winner of this competitive game is the reciprocal strategy widely known as *tit-for-tat*. Too trusting, cooperative and altruistic individuals lose the game because they are too costly, and their traits do not get selected.

investigated incentives for tolerance, disincentives for violence, and occasions of encounter between groups of conspecifics in primates, including early humans. Against the idea that human psychology is essentially biased towards in-group favoritism and out-group discrimination (and even hostility), these data suggest that highly cooperative and tolerant attitudes and behaviors towards strangers were already present in early human societies, as well as in the social behavior of our closest primate relatives (Pisor & Surbeck 2019).

As Kim Sterelny observes,

> (i) there is no clear signature of collective violence on human remains until the very end of the Pleistocene. (ii) There are no battle or raiding scenes in Pleistocene cave art. (iii) The development of projectile weapons, and the associated abilities to stalk and to lie in ambush make chimp-style raids and patrols dangerous, in ways it is not for chimps. (iv) In contrast to skeletal evidence from early farmers, there is little evidence that Pleistocene foragers were routinely resource stressed; that they lived in regimes of near-starvation that compelled a struggle for resources. (v) Jointly, these facts about risks and resource availability suggest that the cost/benefit balance would often favour fairly peaceable relations with neighbours. That is especially true because there are positive benefits from peace […] forager ethnography certainly shows that while foragers are capable of war, they are also capable of peace (Sterelny 2019, 187; see also Sterelny 2016).

This should not, of course, lead us to conclude that biased traits such as parochial altruism and intergroup hostility were not present in early human societies (they were, and they still are present nowadays).[71] However, the fact that significant levels of tolerance and cooperation between strangers have been also present for millennia in human societies suggests that human cognition should not be seen as so rigidly bound to the bellicose and tribal destiny as many have recently suggested in light of ad-hoc descriptions of ancestral environments. On the contrary, a more complex picture of intergroup interactions in early humans and other primates suggests that intergroup cognition and behavior can be significantly more flexible than what the hard-wiring thesis states (on this point, see also Buchanan 2020). Recent research in evolutionary anthropology challenges the idea that parochial and exclusive psychological traits dominated the history of human societies for millennia, showing that also more prosocial, tolerant, inclusive cooperative traits could have, somehow, evolved. But how, and why?

[71] Sterelny notes that "while sceptical of the view that hostility was the default […] in the Pleistocene, the distinction between one-of-us and not one-of-us certainly mattered enormously. I would be astounded if there was anything like an attitude of default trust towards members of out-groups" (2019, 187). However, note that while life in the Pleistocene was probably not dominated by starvation and inter-group conflict, according to forager ethnographic records, hostility and conflict *within* groups were likely pretty high (Boehm 2012).

iii. In the previous sections we considered a core claim usually defended by advocates of the hard-wiring thesis. While, on the one hand, limited prosocial (and even hostile) cognitive and behavioral traits have been selected because they were adaptive (i.e. fitness-enhancing because efficient and functional in facing existential threats in the EEA), on the other hand more inclusive, tolerant, reflective, far-sighted, and open-minded traits have been, comparatively, much more costly and inefficient across human evolutionary history (Haselton & Nettle 2006). However, this claim also seems to contrast with available evidence and theoretical models in evolutionary anthropology. For instance, paleoanthropological data report that tolerance, cooperation and inclusiveness beyond one's narrow social group were actually present in the EEA, and empirical research shows that these traits are even more massively present nowadays (though, of course, not universal). Might they have technically 'evolved' as well? If so, how can we explain the presence of costly cooperative, inclusivist and reflective traits in human interactions, deliberation and social institutions?

We saw above that traits can evolve – i.e. change and spread in populations of organisms – thanks to different evolutionary mechanisms, such as adaptations and exaptations. We also saw that the idea according to which a specific cognitive system dedicated to morality is an adaptation faces several difficulties, and should probably be discarded. Nonetheless, we can hypothesize that reflective and prosocial traits that are not specific to the moral domain evolved in adaptationist terms. Traits favoring more inclusive prosocial dispositions, trust and impartial concern for the interests of out-group subjects might have been selected because they directly conferred several advantages for individuals and groups displaying them: for example, increased opportunities to benefit from broader systems of direct or indirect reciprocity, for the sharing of vital material and epistemic resources, giving birth to stronger offspring as a result of intergroup mating, and so forth.

As Sauer notes,

> a cultural-learning account of evolution can explain why in a highly cooperative niche (i.e. increasingly large communities such as cities), there are adaptive advantages to people having a non-discriminating disposition to cooperate even with strangers and outgroup members (because of rewards from trade or information transmission). The next generation of norm learners […] then acquires this set of undiscriminating norms of cooperation from the previous generation, such that at the end of this process, we have people acting on the simple norm of "be nice to people in general (Sauer 2023, 4.6; see also Sterelny 2019).

Paleoanthropological evidence reports that – as their nonhuman relatives and ancestors – early humans started to associate with out-group strangers very early in the history of our

species. Mostly moved by incentives to obtain useful information – e.g. to cope with common environmental challenges – intergroup contacts increased the risk of threats and danger, but also widely contributed to enhancing reproductive fitness for those people who openly and peacefully engaged in new social exchanges (for example, as a consequence of increased knowledge about how to obtain or produce food, tools, and shelters; see Pisor & Surbeck 2019). Hence, cooperative and tolerant traits in intergroup social exchange could have also been selected because they turned out to be highly beneficial in contexts where greater division of labor between neighboring communities was required (Krebs 2011, 180; Tomasello 2016; see also Buchanan 2020, 152-153).

Furthermore, because of increased chances of access to highly beneficial resources that intergroup interactions created, humans having several fruitful relationships with out-group members significantly increased their reputation and status within their own social groups. Extra-community ties, in fact, can have (had) considerable beneficial effects not only for the individuals directly involved, but also for other members of their community and of the group as a whole. Once again, this phenomenon is mostly due to the acquisition of cost-effective skills and new empowering opportunities and forms of life that intergroup and intercultural contact and exchange can favor (a phenomenon also well-documented by sociological research in post-industrial societies; see Granovetter 1973). As we will see in more detail in the next sections, as far as psychological change is concerned, the explanatory power of recent cultural evolutionary models appears much stronger than that of classic evolutionary psychological functionalist-adaptationist hypotheses about the origins of morality and moral cognition. While giving them too much credit we could be brought to accept the idea of a rigid, genetically predetermined, and universal moral-cognitive architecture, cultural evolutionary models appear much more able to account for massive psychological variations across societies and time, providing both empirical evidence and sound theoretical foundations for the extraordinary plasticity of human cognition and for the flexibility and open-endedness of human morality.

6. Historical and cross-cultural psychological variation

1. Psychology as a historical science

In the previous sections, I argued that the idea that the evolutionary history of our species selected a cognitive infrastructure that is *i)* specifically designed for moral cognition and morality and *ii)* rigidly biased (e.g. towards parochialism and exclusivity) is problematic. Growing bodies of anthropological evidence and recent evolutionary models suggest that the main etiological, functionalist-adaptationist explanations of a hard-wired moral psychology are empirically inaccurate and epistemologically weak. As we have seen above, the social, epistemic, and environmental conditions of the EEA were more complex than the oversimplified picture that can be commonly found in the literature. Moreover, classic evolutionary psychology's core idea of conceiving the EEA-Pleistocene as a privileged time frame in human evolution to which we should refer to explain current cognitive and behavioral traits appears unjustified. If the main reason beyond this view is that the late Pleistocene is the epoch in which our species reached its current anatomical configuration, this view relies on a form of neuro-anatomical reductionism and essentialism about human nature which is frankly untenable, above all because of its blatant contrast with historical and empirical evidence.

The idea that limited cognitive and behavioral traits, selected to address specific ecological challenges in the Pleistocene, would still be widely present nowadays because they are deeply rooted in the human brain and insensitive to socio-cultural interventions and experiences seems, therefore, scientifically poorly grounded. After having considered some empirical and

theoretical shortcomings of the hard-wiring thesis and of the assumptions and methods it relies on, I will now consider some relevant pieces of counterevidence pointing to an opposing position.

In this respect, cultural evolutionary approaches to the study of human cognition and behavior are able to account for significant psychological variations across centuries and different socio-cultural contexts (both over centuries, between generations and within a single life). It is important to keep in mind that the link between socio-ecological enabling conditions and psychological and value shifts are not deterministic, but scientific research can help identify associations and paths that are more likely than others. An extraordinary and still growing corpus of cross-cultural data shows massive variations in people's psychological traits and moral values in different socio-cultural contexts, also providing evidence and insights for understanding the drivers and dynamics which underlie them.

What emerges from the growing body of cross-cultural empirical data about variations in human psychological traits and moral values collected in recent years is that there seems to be basically no universally shared moral cognition, since there is, in the first place, no universally shared human psychology or 'nature' (Prinz 2012). On the contrary, recent cultural evolutionary approaches stress the importance of conceiving of human psychology in more historical,[72] contextual, and anti-essentialist terms, since human cognitive and behavioral traits deeply depend on the society and culture in which they develop (Muthukrishna et al. 2021; Henrich 2020).[73] Since human moral psychology, as we have seen, seems to involve the combination of several multi-purpose cognitive traits, I suggest the same cultural evolutionary approach should be adopted to understand the ecological, social, and epistemic conditions which can enable (or prevent) both the evolution of moral systems as well as collective and individual psychological change.

Across the world, people living in different societies differ considerably in their cognitive and behavioral traits. People in diverse societies – both diachronically and synchronically – perceive, think and behave differently in terms of domain-general reasoning abilities, prosociality, trust and conceptions of fairness (Henrich et al. 2001; Henrich 2020; Muthukrishna et al. 2020; Pinker 2011; Santos et al. 2017), moral judgments (Awad et al. 2019;

[72] To adopt a historical approach does not mean to fully abandon functionalist explanations and, vice versa, also classic functionalist-adaptationist etiological explanations are, in a way, 'historical'. A convincing synthesis between functionalist explanations of psychological-behavioral traits and the significant influence of culture – which can shape human cognition in much shorter times than classic evolutionary psychology would be keen to accept – can be found in Godfrey-Smith's *modern history*'s theory of functions: "the approach is historical because to ascribe a function is to make a claim about the past, but the relevant past is the recent past; modern history rather than ancient" (1994, 344).

[73] According to Henrich, "You can't separate 'culture' from 'psychology' or 'psychology' from 'biology', because culture physically rewires our brains and thereby shapes how we think" (2020, 16-17).

Barrett et al. 2016; McNamara et al. 2019), moral emotions (Elison et al. 2005; Fessler 2004; Kumar & Campbell 2022; Prinz 2007; 2014; Wallbott & Scherer 1995), personality traits (Gurven et al. 2013; Smaldino et al. 2019), beliefs and behaviors concerning personal identity (Ma & Schoeneman 1997), and much more.[74]

Empirical research in cultural psychology and anthropology allows us to identify several conditions predicting significant psychological and value shifts that go far beyond the hard-wired tribal and shortsighted psychology that, according to classical EP and several contemporary commentators (Sauer 2019; 2023; Persson & Savulescu 2012; 2017), humans still carry around since the Pleistocene. This body of evidence clearly shows that not only the *form* and *content* of human moralities (both in theory, social practices, and institutions), but also human *psychology* – the way people think, perceive, and behave more broadly – can change substantially in light of different socio-ecological circumstances.[75] This evidence also strongly challenges the idea that moral change is only a matter of structural-institutional evolutions which leverage on, or totally bypass, a tribal, exclusive and myopic psychology that would be in great part genetically determined and that rigidly remains hard-wired in human brains. Cross-cultural evidence about the relations between different socio-cultural circumstances and psychological variations is massive, and what I report here does not aim at being an exhaustive review of it. Nonetheless, I hope that even a limited set of examples can make clear that significant change beyond psychological tribalism, exclusiveness and myopia is empirically possible and widely documented by scientific data. Hopefully, better knowledge of the dynamics and enabling conditions of these changes will inform future moral and social change theories and strategies in a more realistic and constructive way.[76]

For reasons of space, I will focus here on two main indicators of morally relevant psychological and behavioral variations: the increase and spread of *impersonal prosociality,* and the rise of *emancipative values*. These notions refer to packages of psychological and behavioral traits roughly comprising, or strongly correlating with, among many, the following:

[74] Different socio-cultural environments correlate to different psychological traits, skills and dispositions also in several domains which are not particularly morally relevant. See Henrich (2020, 38–41, 52–55).
[75] This is likely the main difference between my view and Buchanan (2020).
[76] Let me emphasize once again that what I am going to report should not be understood as having intrinsic or direct ethical implications. Nonetheless, this does not mean that such evidence *cannot* have ethical implications. Reflecting on these implications, however, is not my main concern here.

• Increased levels of trust, fairness, cooperation, generosity, and honesty towards strangers, anonymous people and institutions (including impersonal ones, e.g., government);

• Reduced parochialism, in-group favoritism and loyalty;

• The endorsement of impartial, universalist, egalitarian and democratic moral principles, norms and institutions;

• Preference for emancipative values over patriarchal ones;

• Lower levels of conformity and deference to tradition and authority;

• Higher levels of epistemic and behavioral self-control and regulation;

• Desire and love for decisional control and freedom of choice;

• Belief in free will and progress;

• Openness to change and new ideas;

• Analytical thinking;

• …[77]

Empirical research reports significant differences in these indicators over time and across societies (Bond & Smith 1996; Henrich 2020; Inglehart 2018; Welzel 2013). Specifically, levels of impersonal prosociality and the endorsement of emancipative values are much higher in WEIRD societies according to several measures. WEIRD is a now classical acronym in the

[77] I assume that all these psychological traits can be conceived of as morally relevant in the sense that they matter for existing moral practices and challenges, and for our reflective understanding of them. Some of these traits entail or relate to actual levels of altruistic motivation and behavior (e.g. dispositions to donate or to give up one's resources), which can be measurable by means of several paradigms in experimental economics and psychology. Other traits concern moral judgments more directly (e.g. what people judge more explicitly to be right or wrong). Some traits can involve both these dimensions. *Impersonal trust*, for instance, can be assessed both via self-reported answers (e.g. to the Generalized Trust Question: "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?") and via classic experiments with economic games involving real money. Meta-analyses of experimental research on trust suggest that data from economic games and surveys involving the GTQ converge (see Fehr et al. 2002; Johnson & Mislin 2011). For a more complete picture of the many other cognitive and behavioral traits associated with impersonal prosociality, see Henrich (2020, 56).

anthropological and psychological literature – meaning *western, educated, industrialized, rich* and *democratic* – first proposed by Henrich and colleagues (2010). As Henrich repeatedly emphasizes in his work, from a global-historical point of view, WEIRD societies and psychology constitute a very small minority, often placed at the extremes of the global distribution of psychological and behavioral traits (see Henrich 2020, 156-7). Cognitive and behavioral differences listed above correlate with diverse cultural and institutional backgrounds. This evidence supports the high flexibility of human psychology, and highlights conditions and dynamics underlying the directions and shapes that it can assume. I will now briefly zoom into some data and recent hypotheses aimed at explaining these psychological variations.

Note, before, that a similar proposal aimed at understanding the fundamental dynamics regulating moral evolution from limited prosociality and parochial moralities to more complex, consistent, and inclusive moral systems – and even towards more "objective, impersonal, or agent-neutral reasons for action over subjective, personal, or agent relative reasons" (Jamieson 2002, 174) is not new. In the field of naturalistic moral philosophy, a similar proposal had partially already been explored by Peter Singer in *The Expanding Circle* (1981/2011). However, this and other 20th century theoretical attempts to account for significant moral change of this kind relied on much less rigorous scientific methodologies and incomparable empirical evidence at disposal. Moreover, especially in the domain of ethics, theories of moral change too often have combined the descriptive-explanatory analysis of moral change with meta-ethical and normative considerations, inevitably muddying the waters of a scientific understanding of the real dynamics and conditions for moral change.

An exhaustive presentation and discussion of recent evidence about morally-relevant psychological variations in different socio-ecological conditions would require more space. In what follows, I will present and discuss only some of the main drivers that, according to recent evidence, shifted the psychology of people living in certain socio-ecological conditions from being conformist, tribalistic, exclusive and myopic towards significantly higher levels of prosocial behavior, decisional autonomy and farsightedness, and other radical shifts in their psychology and moral values.

2. Evidence and explanations of robust psychological moral change

Disruption of kin-based institutions. Let us begin with a basic fact: our minds and behavior

are to a significant extent influenced by the familiar contexts in which we grow up. Family is the first social reality that humans face once they arrive in the world, and in the vast majority of human societies, it still constitutes the most fundamental institution regulating people's lives, shaping their cognitive and moral development (Sterelny 2010) and their values.

In contexts where social life is rigidly organized around intensive kin-based institutions, people's psychology – beliefs, motivations, emotions, perception – is shaped by the demands imposed by these strong relational ties. Norms of direct reciprocity and care, loyalty, obedience, respect for tradition and authority predominate; freedom, autonomy, openness to change and to new experiences, trust in strangers and the creation of relational bonds outside one's close community and culture are discouraged and much more costly.

Historical, anthropological and psychological data analyzed by Henrich and colleagues in the past few years document that psychological traits such as analytic skills, individualism, independence and impersonal prosociality increase proportionally when the strength of intensive kin-based bonds and institutions decreases. Schulz and colleagues (2019) recently advanced the following hypothesis to explain the emergence of the unique psychological traits that we can find especially in modern and contemporary WEIRD societies: by forbidding marriages between relatives (up to sixth cousins), the multiple bans and prescriptions that the Catholic Church of Rome started to impose since the Middle Ages to regulate sexual and family norms had the effect of disrupting the fundamental social institution of enlarged families and clans which had prevailed basically anywhere in Europe – as anywhere else on Earth – for millennia until then, and that still constitutes the norm for the vast majority of contemporary human societies nowadays (Henrich 2020; Schulz et al. 2019).

One of the main effects produced by these family policies – involving formal prohibitions and sanctions, moral duties and taboos – was the production of an unprecedented *relational mobility* in European regions that were more exposed to the influence of the Western Church. Under those bans, people were forced to create new significant relationships outside their families and close-knit communities: this contributed to the creation of new social institutions and voluntary associations, such as universities and markets. The analyses conducted by Schulz and colleagues report that longer exposure of populations to the Catholic Church predicts *i)* a greater dismantlement of intensive kinship bonds and *ii)* WEIRD psychological variations such as increased impersonal prosociality, individualism, decisional autonomy, and lower propensity to conformism and deference to authority (Henrich 2020; Schulz et al. 2019).[78]

[78] For reasons of space, I deliberately choose not to consider here other fundamental aspects of *religiosity* as drivers of prosociality: see in this respect Norenzayan et al. (2016), Shariff & Norenzayan (2007; 2011).

Recent evidence also shows that weaker kin-networks positively correlate with more participated and democratic institutions (Schulz 2022).[79]

Market integration. A second important corpus of empirical data collected by Henrich and collaborators starting from the early 2000s attests a significant correlation between experiences with (or even mere exposure to) market institutions and the development of higher levels of impersonal prosociality and inclusiveness (Henrich 2000; Henrich et al. 2004). Henrich and colleagues' groundbreaking cross-cultural studies in behavioral economics report that different cultural and institutional contexts significantly shape people's prosociality and other morally relevant cognitive and behavioral traits (e.g. trust, altruistic behavior, conceptions of fairness). In cross-cultural experiments involving the Ultimatum Game[80], people from WEIRD and other industrialized societies are significantly more generous and disposed to accept more egalitarian offers than people living in social realities organized around kin-based institutions, such as hunter-gatherers, herders and subsistence farmers from more than 20 different societies worldwide. The less market-integrated societies are, the more the economic and altruistic behavior of their people reflects the standards of rationality predicted by neoclassical economic models (i.e. maximizing individual utility). On the contrary, people from WEIRD and other market-integrated societies systematically violate these standards (Ensminger & Henrich 2014). By lacking institutions for the regulation of larger cooperative enterprises and challenges, people from less market-integrated societies are psychologically more egoistic (see also Knafo et al. 2009; Marlowe et al. 2008).

Notice, however, that these data do not show that higher levels of market exposure or integration produce mere self-interested and instrumental incentives to increased impersonal prosociality, such as "If I am honest and generous, I will sell more". Even if conditional incentives can certainly play a role in this respect, experimental evidence shows that higher levels of impersonal prosociality typically prescribed by market norms (such as being fair and

[79] A further relevant body of evidence that, for reasons of space, I will not discuss here, concerns the influence of residential mobility on psychological and value change. As research in this field shows, greater residential mobility enhances people's trust in strangers and preference toward egalitarian norms, makes people more tolerant and less exclusive, pushes individual to the creation of larger social networks, to be more open to novelty and new experiences, and even to engage in more creative thinking (Lun et al. 2012; Choi & Oishi 2020).

[80] In the UG, one player (proposer) is given a certain amount of money and has to propose how to divide it between themselves and another player (respondent). The respondent can either accept or reject the offer: if respondent accepts, the amount is distributed as proposed; if respondent rejects, both players get nothing. UG has been a very influential paradigm in experimental economics, and field experiments involving it have suggested relevant insights for the descriptive study of morality. Above all, these and other experimental paradigms in behavioral economics (e.g. Dictator Game, Public Goods Game) can provide information about peoples' sense of (un)fairness, in this specific case by revealing their willingness to punish offers they consider unfair (i.e. deviations from norms of fairness). As shown by seminal cross-cultural studies by Henrich and colleagues, people's sense of fairness and dispositions to punish offers perceived as unfair vary even radically between societies and cultures.

honest with anonymous people), can become so internalized that people also display them in anonymous, one-shot games – i.e. with people they would likely never meet again and who will hardly be able to punish them in the future, even indirectly (Rand et al. 2014). Also, experimental evidence shows that higher levels of impersonal prosociality correlate with higher propensity to engage in voluntary cooperative associations, and with the realization of long-lasting effective formal institutions (Rustagi et al. 2010). This suggests that differences in psychological traits can also explain differences in social institutions, and not only the other way around (more on this below).

Of course, these data do not allow us to infer that mere proximity to market institutions predicts unconditional or purely disinterested moral motivation. However, what matters here is the relevance this evidence can have for the (im)plausibility of the hard-wiring thesis (and, as the reader might have noticed, for the possibility of increased levels of agency and decisional autonomy). Unlike 'institutional bypassing' views (see Sauer 2019, and section 4.4 above), experimental evidence shows that market institutions do not simply *bypass* human psychology and produce positive outcomes as side effects, but they can also radically *modify* human psychological and behavioral traits, not only increasing human agency and decisional autonomy (Welzel 2013), but also expanding the circle of human cooperation, inclusiveness and moral concern (Buchanan 2020; Henrich 2020, chapters 6 and 9).

This conclusion could sound problematic, and a few extra considerations may be needed. It is, in fact, not hard to encounter in several contexts (e.g. in folk morality and academic reflection, both in and outside WEIRD societies) negative moral evaluations of individualist, competitive and calculative traits which are typically seen as emblematic of contemporary WEIRD societies. Surely, these and other WEIRD psychological traits[81] can be problematized from a moral point of view (though this is not my main concern here). However, even though self-interest and competition, for example, are highly valued in WEIRD societies, empirical evidence reports that people from more market-integrated and industrialized contexts also display higher preferences towards zero-sum gains in competitive scenarios when they are obtained by respecting norms of honesty, transparency and fairness, and devalue them when

[81] As well as, of course, market institutions and economies. Deirdre McCloskey observes that "Richer and more urban people, contrary to what the magazines of opinion sometimes suggest, are less materialistic, less violent, less superficial than poor and rural people. Because people in capitalist countries already possess the material, they are less attached to their possessions than people in poor countries. And because they have more to lose from a society of violence, they resist it" (2010, 26). While this may be true, markets and capitalism do not have only 'positive' effects on the psychology, motivations, and behavior of people in societies in which they are more present, since they usually also create and reinforce oppressive hierarchies and inequalities in terms of wealth, health, and power, moral consideration and respect (see e.g. Buchanan 2020, 150-151; Kumar & Campbell 2022, 169; Gowdy 1999).

they are facilitated by forms of partiality and favoritism (Henrich 2020, 294).[82] Although it may sound paradoxical, individualism and altruism are not, in fact, conflicting psychological traits.[83] On the contrary, empirical research indicates that higher levels of performance on individualism scales positively correlate with greater impersonal prosociality and non-reciprocal forms of altruism (ibid.; Rhoads et al. 2021).[84]

It is also important not to misunderstand this evidence with the idea that affluence predicts WEIRD traits such as greater impersonal prosociality, emancipative values, analytic thinking, and so forth. As Henrich notes, historical data clearly indicate that increased levels of income and material security have likely been a consequence, rather than a cause, of the radical psychological and cultural variations and mobility that followed the disruption of intensive kin-based institutions in Medieval Europe. Moreover, evidence indicates that the occurrence of variations towards WEIRD traits is consistent among both wealthy and poor people, suggesting that affluence alone plays no relevant role in shaping those psychological traits (Henrich 2020, 478-480).

Existential security. In his latest work on cultural evolution – collecting more than fifty years of longitudinal and cross-cultural sociological research on value change from more than one hundred countries – Ronald Inglehart (2018) concludes that the massive psychological, behavioral and value shifts experienced by people living in several societies can be primarily explained by a progressive increase in levels of existential security in recent history. According to Inglehart, traits like intensive kinship, strong in-group solidarity, distrust and hostility towards out-groups, moral, religious and political intolerance, and support for authoritarianism are all responses to the precariousness of survival (see also Jost et al. 2003; Tropp 2012, 116). If basic resources are scarce, the risk that there might be just enough either for my tribe or for yours can be high. But prosperity amplifies opportunities of interaction and cooperation,

[82] In this respect, it is worth noting that advocates of the hard-wiring thesis like Perrson and Savulescu (2017, 290) consider *nepotism* as one of the main examples of a hard-wired psychological trait – a claim that appears questioned by this and other empirical evidence showing significant cross-cultural variations in preferences for impartial procedures and norms of fairness.
[83] Individualism should not be confused with self-interest or egoism. Technically, in the psychological, sociological and anthropological literature, individualistic values emphasize the importance of individual autonomy, projects and self-expression over conformity, the protection of one's community and tradition, and the respect of duties, roles and identities which are rigidly defined by them (Henrich 2020; Hofstede 2003; Triandis 1995; Welzel 2013, chapter 6).
[84] I must emphasize, once again, that this evidence should not be understood as something like a 'praise for the West', nor as a defense of the idea that Western cultures or countries are more 'moral' than others. What I have been reporting consists of sets of correlations between socio-cultural ecological variables and psychological constructs; there is no philosophy of history here, destiny, or superiority of the West. As Sauer rightly observes, these "are not Western values any more than mathematics is, in any interesting sense, Arabic simply because that's where the numbers people do it in originated […]. Certain scientific discoveries were, as it happens, first made in certain places rather than others. But these scientific discoveries are part of the universal heritage of humanity. They had to emerge somewhere, but they belong to everyone" (2023, 2.4; see also Welzel 2013, 41-42). Incidentally, empirical research shows that Western values and traditions alone do not predict increases in emancipative values nor in impersonal prosociality (see Welzel 2013)

favoring both psychological and institutional change. In the 20th century, many societies in the post-war era started to experience unprecedented levels of wealth and peace. Probably for the first time in history, several generations in WEIRD countries grew up taking their survival for granted. According to Inglehart, it is such an increase in existential security that led to the emergence of radically new values, beliefs and motivations, which in turn contributed to the rising of emancipatory social movements and more inclusive, fair, and democratic institutions (see also Buchanan 2020, chapter 5. Buchanan calls this 'surplus reproductive success' and regards it as one of the main enabling conditions of the 'Great Uncoupling' and the two 'Great Expansions').

These radical changes in recent history concern all the most paradigmatic instances of moral progress mentioned in Part I; above all, the abolition of slavery; the emancipation of women, LGBTQ+ people, migrants, children, and disabled people; the reduction of several forms of discrimination and oppression (based on ethnicity, religion, political or gender identities); the partial abolition of extreme forms of punishment (such as the death penalty); the institution of international norms against military aggressions, neo-colonialism and apartheid; increasing concern for the condition of non-human animals, ecosystems, and future generations.

Shifts in recent history concerning these issues have been assessed by longitudinal and cross-cultural empirical research both with objective (O) and subjective measures (S), e.g., increasing women employment in positions of high social relevance (O) and people's attitudes towards this trend (S); national laws allowing same-sex marriage (O) and individual perceptions/attitudes towards it (S); and so forth. An influential hypothesis introduced by Inglehart understands these recent global trends of cultural change in terms of a more general shift from 'Materialist' to 'Post-Materialist' values (see Inglehart 2018, chapter 2).

Following Inglehart's theory, for instance, even the growing acceptance of gender equality and homosexuality in certain contemporary societies – according to many, one of the most emblematic recent instances of moral progress (see e.g. Kumar & Campbell 2022, 210-214) – can be explained by increasing levels of existential security. In agrarian societies, for example, high fertility rates are encouraged because they provide an important subsistence opportunity in lack of efficient social security systems. According to Inglehart, this is supposed to lead to developing norms that oppress women and stigmatize sexual behaviors which are not aimed at reproduction (see also Buchanan 2020, 124-125: surplus reproductive success leads to the 'Great Uncoupling', i.e., moralities become free from the demands of reproductive fitness).

Inglehart's theory is solid and useful because it allows to explain value change in all directions: not only inclusivist and/or emancipatory shifts, but also contractions of the circle of

cooperation and moral consideration and respect. Inglehart's research shows that exclusivist contractions of these circles are usually caused by environmental or socio-economic crises, which hamper impersonal prosociality and emancipative values, contribute to the rise of stronger in-group favoritism, out-group stigmatization and dehumanization, exacerbate structural inequalities and power relations, and facilitate the rise of authoritarianism, which leads to a reduction in people's freedoms.[85]

Inglehart's conclusions overlap with Buchanan's (2020), and Buchanan and Powell's (2018) theory of moral change on several points. What emerges from both accounts is an extremely plastic and flexible view of human psychology, since empirical evidence shows that human cognition, motivation and behavior can be either exclusivist and hostile, or inclusivist and cooperative depending on the circumstances. Tribal, shortsighted and hostile responses are more likely to be activated when cues associated with existential threats, such as competition for scarce resources or the risk of contracting diseases, are detected (Buchanan & Powell 2018, 189). When these conditions are not or less present, however, tribalistic attitudes are reduced, and more prosocial and farsighted capacities and behaviors, as well as openness to new ideas and possibilities, become easier to implement (ibid.; Inglehart 2018).[86]

This view portrays a picture of the human mind that also coheres with evolutionary neurobiological models which emphasize brain plasticity. According to these theories, information-processing brain structures are significantly shaped by the environment, and not the predetermined, fixed result of genetic specification. Against the massive modularity hypothesis (see footnote 51 above), neurobiological evidence on brain plasticity suggests that the evolutionary process has not definitively selected/designed specific brain modules rigidly adapted to the ancestral environments of the Pleistocene, but rather selected a brain that is still wide open to change and able to adapt to its environment (Buller 2005, 134-136; Kumar & Campbell 2022; Sterelny 2010).[87] Powell & Buchanan (2016, 246) compare this environment-

[85] Inglehart's own studies show that authoritarian personality traits and values are proportional to the levels of existential security in which people grow up (Inglehart 2018). As in many other theoretical operations – and especially in the domain of social and evolutionary models – clearly identifying causes and effects is a problematic task (technically, a chicken-and-egg problem). The causal links that are suggested here should be understood as orientative, and always conceived of as parts of broader and complex autocatalytic processes. On autocatalytic process and feedback loops between bio-cultural evolution and psychological change, see e.g. Buchanan & Powell (2018, 210-11), Kumar & Campbell (2022, 70-71 and 135-138), Henrich 2015, 57), Sterelny (2012, 29-34).

[86] Although scholars like Inglehart have shown that increased levels of existential security correlate with variations in several psychological traits (e.g. trust, self-control), nobody has proved yet that better material conditions *alone* predict psychological shifts in typically WEIRD cognitive and perceptual abilities (such as analytic thinking, the emphasis on intentionality in moral judgment, the experience of guilt over shame, etc.). On the contrary, the main factors highlighted by Henrich – e.g. variations in family institutions, market integration/exposure, etc. – *predict* several WEIRD psychological variations, highlighting more clearly the causal links between variables.

[87] I want to stress, once again, that this one of the most relevant elements of disagreement between my account and Buchanan (2020). According to Buchanan, moralities *but not moral psychology* (or 'the moral mind') are flexible and not rigidly fixed by evolution. My view is even more radical, and claims that empirical evidence gives us reasons to believe that also moral

dependent plasticity of the human mind to the sophisticated armor that certain water fleas develop only in the case in which they detect chemical elements related to the presence of predators or other existential threats (for further evidence on adaptive phenotypic plasticity, see Stoks et al. 2016; Reger et al. 2018; Watkins 2021). As recent empirical research and cultural evolutionary models suggest, the influence of culture on human psychology, behavior (and phenotype more generally) is so deep that it can even strongly shape the evolution of human physiology and anatomical structure (Henrich 2015; Sterelny 2012; Wrangham 2009). Therefore, it seems that there is good reason to conceive of human moral psychology and behavior as the extremely plastic and malleable result of several flexible cultural, societal, and cognitive mechanisms, rather than as rigidly bounded by the biological evolution of our species, genetically entrenched and hard-wired in human brains.

3. Limits and critical considerations: What about agency?

The evidence reported in the previous sections challenges the hard-wiring thesis and its evolutionarily explanations. On the one hand, the idea that tribal, biased, and exclusivist traits have been selected in the Pleistocene and still strongly influence human behavior contrasts with evidence and theoretical advancements in the psychological, sociological and anthropological research. On the other hand, recent longitudinal and cross-cultural studies report that massive psychological and value variations can occur even in relatively short timespans, portraying human cognition, values, and morality as extremely plastic and malleable (see for instance Lun et al. 2012; Choi & Oishi 2020).

Nonetheless, this evidence also depicts psychological and moral change as highly dependent on macro-level socio-environmental conditions, whose dynamics are often hard and costly for individuals to understand and control. Buchanan and Powell's naturalistic theory of moral change (2016; 2018)[88] emblematically represents this view: in their account, inclusivist and emancipatory shifts can only occur under favorable socio-ecological conditions. In their own words, moral inclusiveness and emancipative values are 'luxury goods' (2018, 210–17; see also Buchanan 2020).

psychology is much more flexible than many have recently claimed. Thanks to Allen Buchanan for pushing me to highlight this point.

[88] Buchanan and Powell developed a comprehensive theory of moral *progress*. However, as stated above, I am mostly interested in their underlying descriptive-explanatory theory of moral change. I will thus focus on this latter by conceiving the moral shifts they deem progressive from an empirical, naturalized perspective, i.e. leaving aside evaluative normative and metaethical considerations about why they are good or desirable instances of moral change. In other words, I will mostly focus on why and how some of the kinds of moral change they have in mind occur, i.e. their enabling conditions.

I concur with most of Buchanan and Powell's theory. It should be stressed that, among contemporary accounts of moral change/progress, their account is one of the few emphasizing the importance of the psychological dimensions of moral change and acknowledging the scientific plausibility of significant moral and cognitive plasticity and flexibility. As we saw in Part I, despite some theoretical limitations on the normative-evaluative level, their view seems particularly concerned with finding a place for human agency in the dynamics of broader socio-institutional moral shifts (Buchanan & Powell 2018; Sauer et al. 2021).[89]

However, their theory appears unable to provide a sufficiently convincing *naturalistic* account of the enabling conditions for the increase and exercise of agency and open-ended-normativity, as well as of their causal role in the promotion of macro-level moral change (if agency-driven moral change is the most important kind of progress, how is it possible?). The critical considerations I will now propose against Buchanan and Powell also apply to other accounts of moral change which are more explicitly materialist, supra-individualist, and/or structuralist-institutionalist in nature (like Sauer 2023). These accounts, in fact, are even less able to make sense of significant psychological change and increase in the causal role of agency and open-ended reasoning and normativity in psychological shifts (see Klenk & Sauer 2021), cultural evolution, and institutional change. As anticipated above, for some of these views this is not much of an issue: they are basically eliminativists about these factors, believing that significant moral change is easily explainable without referring to them (Sauer 2019) and, as we saw in Part I, unimportant to assess the progressive moral value of societal change.

Above all, Buchanan and Powell's view seems to face a methodological-explanatory dilemma:

(i) First, they could admit endorsing a materialist, structuralist-institutionalist view, attributing a strong causal priority to factors and dynamics which are mostly beyond agential control, significantly downsizing the causal role of agency in the explanation of the instances of moral change they are mostly concerned with. By doing this, however, they expose their theory to critiques of materialism or environmental/socio-structural determinism, conservatism, and lack of helpful action-guiding criteria or principles to orient social change and/or moral reform. Buchanan and Powell are ambiguous on this point: they explicitly declare not to endorse such a view, but their theory often seems to point in that direction.

(ii) If, alternatively, they claim to attribute a relevant *causal* role to factors such as human

[89] For another view significantly concerned with moral/cognitive plasticity and the psychological dimension of moral change, see Kumar & Campbell (2022).

agency, open-ended moral reasoning and normativity in the dynamics of moral change – as their account of moral progress would necessarily require (see Part I, 3.2 above) – then their theory seems unable to account for both the enabling conditions and the causal role of these factors on broader societal shifts in fully naturalistic terms.

Buchanan & Powell provide no theory nor evidence to suggest how these capacities can emerge and develop; how their exercise and improvement could be possible and promoted; and how these "moral powers" (2018, 50) they refer to play a causal role in the broader societal shifts they discuss. In either case, their view does not respect the main naturalistic requirement of the meta-theoretical framework that I proposed in chapter 1, i.e., the provision of a realistic descriptive theory of moral change. Recall that such a theory is crucial in a theory of moral progress that aims to give a few insights on how to avoid regress or foster progressive shifts. Because of this limitation, Buchanan & Powell's account also risks being even less helpful for both theoretical and practical aims (but see Buchanan 2020 for a more convincing view in this respect). A closer analysis of the two horns of this dilemma will show that neither of the two constitutes a satisfactory naturalistic account of moral change.[90]

Horn (i)

On the one hand, if a theory of moral change like Buchanan & Powell's emphasized such a strong one-way causal relation of dependence from macro-level socio-environmental conditions to micro-level psychological and value change – without giving much credit to other factors (i.e. agency, conscious reasoning) – it risks exposing itself to accuses of materialism or environmental/socio-structural determinism, i.e. the view according to which "ultimately, broadly 'material' forces are responsible for driving people's values or, at the very least, for

[90] Buchanan (2020, 139-143) tries to correct this shortcoming of Buchanan & Powell (2018) by providing i) an evolutionary explanation of moral consistency reasoning (MCR; see Kumar & Campbell 2012) based on a 'partner choice' hypothesis, and ii) by emphasizing the fundamental motivational role of moral identity to engage in MCR. However, both MCR and moral identity are conceptually different from the ideas of agency and open-ended normativity. First, in Buchanan's own view, MCR seems to have a pretty clear evolutionary and social function (that of being seen as reliable and predictable cooperative partners), while agency and open-ended normativity do not seem to necessarily have that function (on the contrary, the specificity of open-ended normativity should be precisely that of being untethered from any kind of functional normativity). Also, agency intrinsically includes motivation while MCR does not; and while agency and open-ended normativity are defined by including counterfactual and prospective thinking, MCR is mostly a kind of systematization of existing beliefs. Finally, MCR seems a very demanding indicator of partner reliability (it is hard to imagine that social actors infer the trustworthiness of their partner from how well they engage in MCR – actually, what counts as good MCR is a matter of dispute even among professional scholars in the field). Second, as fare as moral identity is concerned, it seems to me that moral identity as Buchanan conceives it can basically just make people comply better with new ('progressive') values, but not to engage in complex, critical open-ended moral reasoning. Again, it seems a little too optimistic to think that people want to be seen by others as agents who engage well in sophisticated MCR; this may be true, but just for a very restricted number of individuals.

providing the fertile ground that allow values whose time has come to thrive" (Sauer 2023, 2.4). This view might be not problematic per se (see Sauer 2023), but I am not sure Buchanan and Powell would endorse it. If we agree that supra-individual, socio-environmental conditions and dynamics are not easy to understand and control, and social structures to change, by depicting psychological and value change as so contingent on them such an account risks to provide a view of human psychology, values, behavior and moralities that is hardly susceptible to autonomous and consciously designed projects of reform. But this, on the evaluative level, is the most important kind of moral progress for them: moral progress "in the robust sense" (2018, 52).

One might respond that my concern would be justified only if Buchanan and Powell believed that the socio-environmental enabling conditions they have in mind are not only necessary, but also sufficient for moral change. As a matter of fact, however, Buchanan and Powell seem to suggest this reading in several passages, such as in the flea's armor analogy mentioned above. Other critics have noted that, according to Buchanan and Powell, "Socio-cultural innovations that alleviate infectious disease, resource scarcity, physical insecurity, interethnic conflict and low rates of productivity *suffice* to foster inclusive morality" (Persson & Savulescu 2017, 287; cf. Powell & Buchanan 2016, 247).

Moreover, in this way Buchanan and Powell also seem to concede too much to the pessimist or to the conservative (Sauer 2019), risking making the very proposal of actively designing and struggling to implement strategies for moral and social reform a worthless ideal. In fact, if external conditions are not luxurious enough to allow the required/desired moral shifts, then any effort is basically vain – but harsh social conditions are usually those in which these changes would be most needed. Should people give up until material or institutional conditions improve?

Buchanan & Powell's theory would not collapse into a materialist or structuralist-institutionalist account were it able to explain not only what favors the exercise or improvements of human moral capacities – both in favorable and unfavorable conditions – but also whether and how the agential capacities they refer to can have an impact on broader structural-institutional dimensions of morality. Can psychological change and the exercise of agential capacities contribute to structural-institutional moral change even in harsh socio-ecological conditions?

Horn (ii)

Buchanan & Powell may reject the first horn of the dilemma, i.e. the critique that accuses them of positing too strong a dependence of moral change on environmental-material conditions. They could claim that their theory also wants to account for other factors in the dynamics of moral change, and that material, structural, and environmental conditions are not sufficient to explain important moral shifts, including psychological ones. In fact, parochial, conformist, intolerant, myopic and authoritarian psychological traits are widely present in human populations even under what Buchanan and Powell consider luxurious material and institutional conditions. What happens when favorable material conditions do *not* correlate with increased prosocial, inclusivist, emancipative psychological and value shifts? Buchanan and Powell try to explain this phenomenon by claiming that tribalism and exclusivist traits and institutions can emerge and stabilize even if actual existential threats are absent, but people still *perceive them as real*, for example as the result of ideological or demagogic manipulation (Buchanan & Powell 2018, chapters 6-7)[91,92]. Although this hypothesis is plausible, its general explanatory power remains limited.

While the socio-environmental circumstances discussed in the previous sections certainly play(ed) a relevant role in shaping human psychology towards greater impersonal prosociality and emancipative values and institutions, they still appear insufficient – and perhaps even *unnecessary* – to explain them. Even taking into account phenomena such as ideology and demagogic manipulation, I believe we still need to understand more of the micro-level mechanisms involved in moral change dynamics (both inclusivist and exclusivist, emancipative and oppressive), e.g. why certain individuals and groups are more or less susceptible to them, and under which conditions and how these influences can be avoided and contrasted (see Zmigrod et al. 2021; Zmigrod 2022).

If Buchanan and Powell decided to take the second route – i.e., declared to be open to

[91] Taking into account the subjective perception of one's existential condition – together with more 'objective' measures of it (e.g. well-being, institutional stability) – seems in fact very relevant for understanding people's social and moral psychology. For instance, Rhoads and colleagues (2021) found strong correlations between levels of subjective well-being (self-reported life satisfaction), individualism, self-expression values, and several prosocial behaviors. These data could be explained by the fact that several forms of impersonal prosociality such as non-reciprocal (or non-conditional) forms of altruism are often driven by autonomous decisions rather than being prescribed by the norms and expectations of one's close-knit community (more on this below).

[92] Buchanan and Powell call 'moral regressions' those exclusivist shifts and conservative or reactionary pressures which are directed against expansions of the moral circle or their institutional stabilization (2018, chapter 7). In the beginning of this work, I declared my intention to treat the phenomenon of moral change, as much as possible, in neutral, descriptive terms. I will then stick with the distinction between inclusivism vs. exclusivism, impersonal prosociality vs. intensive kinship (or tribalism), emancipatory vs. patriarchal values, and so forth. I believe that these concepts are much less controversial than the ideas of moral *progress* and *regress* for the naturalistic approach adopted here, since they are 'thicker', their meaning does not depend on substantive normative or metaethical views, and (also for these reasons) they are likely more immediately intelligible by anybody.

consider further explanatory factors – without renouncing their naturalistic methodology, their account of moral change will still appear methodologically and empirically defective, and practically only moderately helpful. In fact, the enabling conditions they focus on appear insufficient to understand retrospectively the dynamics underlying the kind of psychological moral change that constitutes the common object of our works. Even more importantly, their account also appears insufficient to develop prospective empirically-informed and effective strategies to foster (or avoid) further change, something their theory would like to be a guide for (2018, 31): how do we change material conditions, institutions, and social structures if not by engaging directly in reflection, social critique and action, and/or trying to suggest other people different ways to reason and act to promote our common goals?

What else, then, if not only materials conditions can lead people to change the way they reason about morally relevant issues? What is it that drives people to decide instinctively and inflexibly (e.g. under the influence of natural inclinations, habits and/or local socio-cultural norms) or, instead, in light of more careful considerations of options, information, interests and reasons? I believe that failing to consider this question constitutes one of the greatest weaknesses of contemporary accounts of moral change. In what follows, I try to fill this gap.

7. Explaining open-ended normativity

1. Agency and open-ended normativity: evolutionary mysteries?

So far, the main aim of this work has been to show that psychological moral change beyond tribalism, parochialism and shortsightedness has massively occurred across societies, especially in recent history. Empirical evidence indicates that, in certain ecological circumstances, even people's *motivations*, and not only their beliefs – one of the main concerns expressed, e.g. by Persson and Savulescu (2017) – have been changing towards greater levels of moral inclusivity, impersonal prosociality, farsightedness and emancipation. This process is still ongoing, and there seem to be no striking reasons to believe that humanity (who?, where?) has reached a definitive and genetically specified limit to moral inclusiveness, emancipation, and cooperation (for a similar 'fatalistic' view about moral psychology see also Klenk & Sauer 2021; Sauer 2019). The plausibility of the hard-wiring thesis appears, once again, significantly downsized.

Notwithstanding, morality is a much more complex phenomenon than that outlined so far. The phenomenology of moral experience involves more than merely trusting (or not) strangers, avoiding discriminating against them, or refraining from being violent in competition for scarce resources. Some of these (and other) behaviors are nowadays clearly specified and demanded by formal and informal norms and institutions – especially in WEIRD societies – whose foundations and justifications would be likely endorsed by most of the readers of this work.

Some of them are even encouraged by effective systems of sanctions and incentives that – paired with other mechanisms – make them even more easily internalizable, habitual, and widespread in WEIRD societies.

But certain contemporary problems – including the possibility of being critical towards received moral and non-moral principles, norms and judgments – are more controversial and complex than others. Even if oftentimes no unequivocally correct answers can be given to them, some moral problems require careful contextual examination, information-gathering, reasoning and justification. Morality does not merely require *feeling* or *behaving* in specific ways, but it also involves the peculiar activity of critically reflecting, deliberating, and justifying choices, judgments and norms concerning complex interpersonal and/or collective problems (see Songhorian et al. 2022).

Certain social practices, moral norms and principles can be (or become over time) relatively uncontroversial in theory,[93] though they might remain difficult to implement in practice.[94] Others, however, are much more controversial in theory, and this can make them difficult to implement. Cases of the former kind mainly concern motivation; cases of the latter kind, reasoning.[95] The evidence discussed so far indicates that favorable socio-ecological circumstances mostly predict change in people's motivations, dispositions to trust, self-control, and other multi-purpose or domain-general capacities involved in moral experiences, reasoning, and decision-making. In what follows, we will see how and why moral agents can combine these capacities together in different ways that are relevant for the present discourse.[96]

As we saw in the previous sections, according to some evolutionary psychologists it is hard to account for the selection of a costly cognitive capacity to engage in sophisticated moral reasoning, because quicker and cheaper solutions such as emotions and heuristics are often more efficient to address moral problems. Some scholars have suggested, however, that the

[93] More on this in Part III.

[94] Several scholars have suggested that the wider the circle of moral consideration gets, the harder it is to practically satisfy the normative demands that could follow from this extension. See e.g. Asma (2012), Goldsmith & Posner (2005), Haidt (2012), Persson & Savulescu (2017).

[95] Such a theoretical distinction does not imply that reasoning and motivation should be kept distinct: reasoning can, at least in part, contribute to determining judgment and action (see respectively Sauer 2017; May 2013), and non-doxastic motivations are known to influence moral reasoning (Bago & De Neys 2019; Haidt 2001; Kunda 1990; Mercier & Sperber 2017).

[96] This perspective recalls the radical anti-consequentialist and anti-externalist methodology declared at the beginning of this work. Actual behavior and social outcomes should certainly figure among the main objects of analysis in a theory of moral change. However, I believe that a theory of moral change (as a theory of morality more broadly) should also account for variations in 'internal' cognitive processes involved in moral decisions rather than focusing only on observable behavior and the effects that it produces on the world. This thesis might be misunderstood as evaluative, but here I conceive it mostly as methodological. Regardless of which of these aspects matters more from a moral point of view, my position merely acknowledges that each level of analysis can be studied empirically, that all play a relevant causal role on the social reality, and that we dispose of reliable knowledge and power to intervene on each of them. Whether we conceive this methodological starting as an a-priori theoretical assumption or as an empirical observation, I believe such a pluralist methodology gives us back a more complete model to understand the social reality, as well as greater chances to change it (especially, once we understand links and relations between different levels. See Madva 2016).

emergence of such a capacity might be explained by the positive social function of *post-hoc* rationalizing potentially controversial choices of high socio-emotional relevance to make them appear more acceptable to other people (Haidt 2001; 2012; Mercier & Sperber 2017). According to this view, however, this capacity plays no *ex-ante* causal role in the determination of moral judgment and behavior, which are essentially the direct result of emotionally-laden reactions of approval and disapproval (Haidt 2001; 2012). Recently, several scholars have severely criticized this view (see Cushman 2020; Campbell & Kumar 2012; Fine 2006; Railton 2014; Sauer 2012; 2017; Songhorian et al. 2022, section 4; section 11.3 below).

Evidence and models reported in the previous paragraphs do not account for the causal role played by moral reasoning in the dynamics of moral change. Moreover, apart from a few exceptions (Bina et al., ms; Buchanan 2020; Campbell & Kumar 2012; Kumar & Campbell 2022; Huemer 2016; Singer 1981/2011) many recent contributors to the debate on moral progress which take a naturalistic and broad historical and global perspective on moral change tend to reject (or significantly downplay) the causal role of moral reasoning in the promotion of inclusivist moral shifts and emancipative values (Hopster 2019; Rorty 1999; Severini 2020; Smyth 2020; Tam 2019).

As already stressed, data and theories discussed in the previous chapters suggest a strong dependence of psychological and value change on environmental and socio-economic circumstances. Specifically, harsh conditions seem to correlate with tribalism, myopia, exclusivism, deference to tradition and authority (etc.) while prosperity leads to greater inclusiveness, open-mindedness, and cooperation. So conceived, these models leave little room for the possibility of active, intentional and reasoned projects for moral reform. By depicting change in moral beliefs, attitudes, and motivations as so dependent on external circumstances and supra-individual dynamics rather than on human agency, this view seems to weaken the very idea of 'open-ended' moral normativity and its changing potential.[97]

The concept of 'open-ended normativity' indicates the possibility of deviating from the strong external influence of other kinds of normativity human psychology, behavior and

[97] An alternative approach may understand social, cultural and moral changes that appear too 'big' for our limited evolved psychologies as relying on a sort of external 'extension' or 'scaffolding' of human cognition (see e.g. Gallagher 2013; Sauer 2019). To my knowledge, a systematic theory of the 'extended moral mind' has not been proposed yet. But I suspect that, if developed, such a theory would still have to directly deal with the plausibility of the hard-wiring thesis. Either one believes that significant moral shifts (e.g. emancipatory, inclusivist) can only occur 'outside' of individual minds, because of the evolved constraints of human psychology (Sauer 2019; 2023), or we have to acknowledge that these kind of moral change can be both *driven by* and *cause of* substantial individual psychological change. My current research makes me more inclined towards the latter view, which I see as less reductionist and more pluralist on the methodological level. Joe Henrich and colleagues' work is perhaps the strongest evidence in its support.

moralities are often exposed to.[98] This concept is very close to the ideas of agency and decisional autonomy introduced in chapter 3. Buchanan and Powell consider this capacity a necessary condition for the realization of inclusivist shifts, claiming that "any naturalistic explanation of inclusivist morality must feature the capacity for open-ended normativity" (2015, 65). Unfortunately, however, Buchanan and Powell are unable to explain in *even minimally* naturalistic terms why and how this capacity emerged and could be developed, implemented, and improved. The authors limit themselves to assume its existence as a fact, providing only a few anecdotal examples and failing to explain what might favor or hamper it; in their own words, "It is clear that this capacity exists [...] even if we do not possess a good account of the conditions under which the capacity is likely to be effectively exercised" (2015, 64).

Buchanan and Powell are not simply saying that we currently lack convincing explanations for the emergence and development of this capacity. Their much more radical claim is that the ability for open-ended normativity, as the inclusivist moral shifts that according to them necessarily depend on it, *cannot* be afforded by *any* evolutionary explanation – "whether of the selectionist or by-product variety" (2015, 38)[99].

Above I reported some skeptical considerations about the idea that a specific capacity for moral cognition could be considered an adaptation – i.e., that it was selected because of, and still would perform, a clear fitness-enhancing evolutionary function. However, as anticipated, not every evolved trait is an adaptation. Some traits are hardly explainable by referring to the function(s) they perform within a certain ecological system, but their emergence and persistence can still be explained as the scientifically and logically plausible consequence of the selection of other traits, or the combination or co-optation of traits for uses which differ from the original function for which they have been selected.

As introduced above, these derivative traits are known in the literature as evolutionary by-products, a peculiar kind of which are called *exaptation*s (see Gould & Lewontin 1979; Gould & Vrba 1982; Buss et al. 1998). In their critiques of evolutionary explanations of morality, both Machery & Mallon (2010) and Buchanan & Powell (2015) not only reject the idea that morality and moral cognition may be considered evolutionary adaptations, but also that they could be

[98] For instance, according to several evolutionary models, many social norms in virtually any culture are originally selected because of their contribution to enhance biological fitness, both phylogenetically (Haidt 2012; Joyce 2006; Tomasello 2016) and ontogenetically (Railton 2017). More on this topic below.

[99] It is no accident that Buchanan & Powell call these shifts "inclusivist anomalies". The main reasons behind this claim are (a) that adaptationist hypotheses explain the selection of traits by referring to specific evolutionary functions, but the capacity for open-ended moral reasoning doesn't seem to have a clear one;(b) by-product hypotheses are basically just-so-stories (i.e. *ad hoc* conjectures with no sound scientific basis). As I argued in the previous sections, (a) is a plausible claim. However, as I will argue below, (b) is problematic.

understood as the by-products of other evolved traits (Buchanan & Powell 2015, 55-60; Machery & Mallon 2010, 23).

Although they provide several arguments and evidence against the idea that a specifically moral cognitive capacity technically 'evolved' (some of which I have reported above), Machery and Mallon provide no argument in favor of their skepticism about conceiving moral cognition as a by-product of other evolved traits. On the contrary, they suggest that

> the capacity to grasp moral norms and the capacity to make moral judgments might be similar to chess or handwriting. The capacities to play chess and to write involve various evolved cognitive traits (e.g. visual recognition and memorization of rules for the former), but they did not evolve. Similarly, we conjecture that the capacity to grasp moral norms and the capacity to make moral judgments involve various evolved cognitive traits (including [...] a disposition to grasp norms in general), but they themselves did not evolve (Machery & Mallon 2010, 23).

In other words, what they are saying is that moral cognition is the consequence of other evolved traits, such as basic building blocks of social cognition in animals (e.g. emotions) or a distinctive capacity for normative cognition. But this resembles what a byproduct explanation of moral cognition would look like. We could agree with Machery and Mallon that moral cognition was not *directly* selected in functionalist-adaptationist terms (i.e. because of its direct contribution to improving reproductive fitness), but to deny that conceiving its emergence as a byproduct of other capacities is also a plausible evolutionary explanation seems to be a conceptual mistake.[100]

Buchanan and Powell (2015) provided slightly more articulated but still inconclusive arguments against by-product explanations of moral cognition, and, in particular, of the capacity for open-ended moral normativity. Their arguments stress that no available by-product explanation is sufficiently rigorous and detailed to conclude that open-ended moral reasoning and normativity are necessary implications of the selection of other traits – e.g., as the spotted hyena's hypertrophic clitoris is a necessary biological consequence of its evolved increased aggressiveness, or as the Spandrels of San Marco (the triangular-ish architectonic elements between the arches) are the secondary but inevitable consequence of the design and

[100] To be honest, even Machery and Mallon's direct critique of adaptationist models of the evolution of morality (2010, 24-30) is not particularly convincing. Their argument goes roughly as follows: models like reciprocal altruism and indirect reciprocity aim to explain the emergence of morality in adaptationist terms. These models do provide plausible explanations for the evolution of certain (limited) forms of reciprocity, prosociality and cooperation, but these mechanisms constitute only a small part of morality. Conclusion: morality is not an adaptation, because these adaptationist models are unable to account for more impersonal and non-reciprocal aspects of morality. However, by no means this argument tells us that these latter forms of prosociality cannot be given *any* evolutionary explanation. It simply suggests that the explanatory power of those specific adaptationist models is limited.

construction of more fundamental structural elements of the building (Buchanan and Powell 2015, 56; Gould and Lewontin 1979).

Here some considerations are needed about what can be technically considered a byproduct, especially in relation to the degree of probability (or absolute necessity) according to which an evolutionary 'spandrel' is supposed to follow from the selection of other traits. If we can call byproducts only *necessary* consequences of other traits, then very likely we do not have sufficient data and instruments to conclude that the psychological traits enabling open-ended normativity are *necessary* consequences of the selection of other traits (e.g. emotions, empathy, theory of mind, domain-general reasoning abilities, or norm cognition). But if this is true, we cannot even claim, as both Machery & Mallon and Buchanan & Powell do, that they are certainly *not* byproducts. The possibility that the cognitive capacities allowing the development and exercise of open-ended moral reasoning and normativity are necessary consequences of the selection of other traits remains, in principle, absolutely plausible (though this does not imply that their implementation, or certain specific outcomes – e.g. specific emancipative or inclusivist shifts – are necessary consequences of the selection of those traits as well).

If, on the other hand – as Buchanan & Powell seem to suggest – we can call byproducts *scientifically and logically plausible* (though not 100% certain) consequences of the selection of other traits, then to affirm that open-ended morality and moral cognition can be understood as by-products sounds even less problematic. Moreover, as we saw in the previous sections, psychological traits can vary fairly rapidly because of the influence of cultural pressures. Both the quantity and variety of possible affecting factors and the speed of cultural evolution make potential byproduct explanations of cognitive traits such as open-ended moral reasoning qualitatively different from the cases of San Marco or the spotted hyena's clitoris. We cannot demand the former the same minimalist and linear causal sequence that is needed to explain the latter.

Although Buchanan and Powell "do not mean to advocate any mysterious or transcendental view" (2015, 65), their rejection of *any possible* evolutionary explanation for moral cognition, open-ended normativity and inclusivist shifts makes their views scientifically suspect, metaphysically problematic, and the characterization of their theory of moral change as 'naturalistic' conceptually incorrect (FitzPatrick 2019). Positing such a gap between scientifically sound evolutionary principles and a peculiar ethical capacity and/or process that *i)* can operate independently/regardless (and even radically against) the former, and *ii)* is unexplainable in scientifically/naturalistically plausible terms recalls problematic dualisms in

the history of evolutionary theory (Huxley 1893)[101], and casts doubt on the very naturalistic methodology Buchanan and Powell declare to rely on in their work (2018, 26-30), moving closer to non-naturalistic accounts of morality and ethics.[102] The limits of these accounts thwart the very project of explaining open-ended moral cognition and robust inclusivist-emancipatory psychological moral shifts in fully naturalistic terms, and hence any ambition to use available scientific knowledge to design empirically-informed strategies for promoting moral reform (or avoiding undesirable shifts).

> Presumably, however, this capacity has been around for quite some time. Why, then, did the dynamic of expansive morality not happen earlier, or indeed later? Why did it happen, when it happened? This is, again, not problem that afflicts their [Buchanan & Powell's] theory specifically; rather, it is a formidable puzzle any theory of moral progress (or regress) must contend with (Sauer 2023, 1.3).

In the next section, I seek to fill this gap by exploring the possibility of explaining humans' capacities for open-ended moral reasoning and normativity as evolutionary byproducts.

2. Enabling conditions for open-ended normativity: a naturalistic, cultural-evolutionary explanation of increases in agency and agency-driven moral change

In what follows, I argue for a more optimistic view about the possibility of explaining agency-based moral change in naturalistic terms.[103] I propose a model that integrates the theories discussed above by stressing and clarifying the links between macro (ecological, structural) and micro (psychological) levels of moral change. I do this by offering a causal sequential model that explains (a) the cultural selection of psychological traits enabling (or hampering) agency and open-ended normativity, and (b) the latter's role in the dynamics conducting to (or hampering) macro-level structural-institutional shifts.[104]

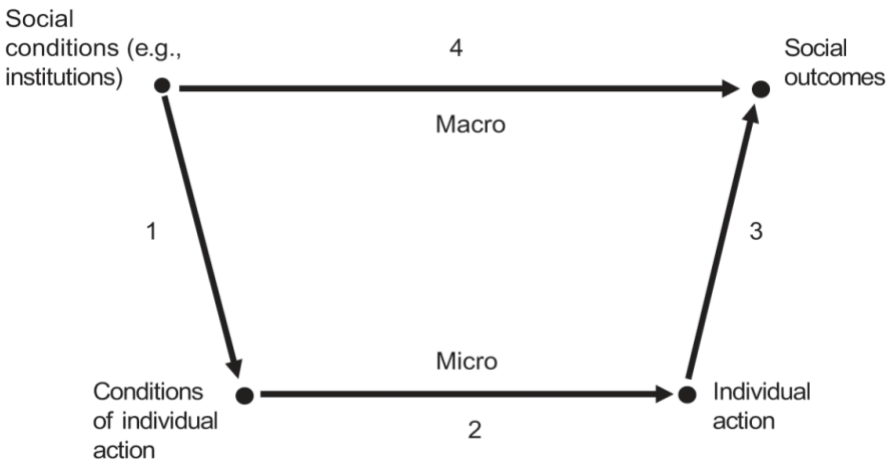
My model rejects views according to which either moral change is *identifiable* with mere structural-institutional change, or it involves psychological change but only as an epiphenomenon of partly ungovernable supra-individual dynamics. On the contrary, according

[101] Notably, a similar paradox had been spotted in Huxley's idea of a fundamental opposition between the natural-cosmic and the ethical process. Huxley's view in *Evolution and Ethics* is synthesized in his famous metaphor depicting the cultural-ethical process as a gardener who controls the spontaneous process of life and death of the vegetation, naturally characterized by competition for resources and prevarication.
[102] See Nagel (2012) for a fairly isolated defense of the thesis that the apparent incompatibility between 'Darwinian' naturalism and our phenomenological experience and understanding of morality, freedom and consciousness casts doubt on the former.
[103] With 'agency-based' I mean that agency is understood together as the source, the end and the product of moral change.
[104] The basic structure of my model draws on a human development sequence proposed by Welzel and Inglehart (2010), but it substantially differs from it for a) its specific focus on psychological change, agency, and open-ended normativity and b) its reliance on a more pluralist and empirically-grounded set of theoretical assumptions concerning human motivation.

to my model, psychological moral change – i.e. change in individuals agency, beliefs, reasoning abilities, attitudes, motivations, and several other cognitive-behavioral traits and skills can be both the product and driver of broader structural-institutional shifts.[105]

Simplifying, we can illustrate the general dynamics of moral change and potential interplays between macro and micro levels of analysis by drawing on Coleman's 'bath-tub' model of social change (Coleman 1990, Fig. 1). In a nutshell, according to this model, changes in macro-level socio-environmental conditions foster changes at the micro level of individual motivations, psychology, decision strategies and action, which in turn concur to promote further change at the macro level of social structures and institutions.

Fig. 1. *The bath-tub model of social change*.



In the previous sections, I reported some preliminary (and certainly partial) evidence in favor of links 1 and 2 – i.e., psychological and value shifts fostered by environmental, structural, cultural and/or institutional change. But what about links 2 and 3 and the node between them, representing individual action? Can psychological and value change allow (or hamper) both the micro-level conditions for greater agency/open-ended moral reasoning and normativity and drive macro-level social and moral change? How?

Before answering these questions, let us briefly wrap up and recall a few relevant theoretical points we touched so far. At the beginning of Part II, I presented and discussed an influential view in the contemporary debate on moral change: the idea that selective pressures in the

[105]According to Inglehart and Welzel, institutional change is often *preceded* and *caused* by psychological and value change, and not vice versa (Inglehart 2018, 3; Welzel 2007; 2013, 38; Welzel & Inglehart 2010).

Pleistocene forged a tribal, myopic, exclusive psychology that is strongly insensitive to change and to 'ordinary' projects of moral reform. Times of biological evolution are very slow and, advocates of this view maintain, psychological evolution works in a similar way.

I then showed (4.4) some implications of accepting this view, by outlining two accounts of moral change that have been proposed in recent years by relying on the hard-wiring thesis. Specifically, one of these views emphasizes that significant psychological change beyond parochialism and myopia, as well as improvements in agency (i.e. better reasoning *and* motivation) is basically impossible to achieve with ordinary socio-cultural means, and we should invest in alternative, more effective strategies such as moral bioenhancement (Persson & Savulescu 2012; 2017). According to the other view, psychological change plays basically no role in the promotion of significant moral shifts, which rather occur outside individual minds, at the macro-level of social structures and institutions, and is even favored by limited prosocial dispositions, self-interest, and desire for power if these are institutionally channeled in the correct way (Sauer 2019).

In chapters 5 and 6 I reported several considerations and lines of evidence showing that the empirical claims these views rely on are fairly weak. In particular, I stressed how massive psychological variations towards greater prosociality, normative emancipation, decisional autonomy, farsightedness, self-control (etc.) occurred throughout human evolution, especially in recent history and even within a short time. The idea that human moral cognition is now rigidly fixed and constrained since the Pleistocene, and only supra-individual dynamics and processes constituted and/or drive significant moral change seems to be empirically unsupported and in contrast with available evidence. But what are the links between 'macro' and 'micro' levels, and how can we account for psychological moral change and increases in agency and decisional autonomy in naturalistic but not deterministic terms?

Returning to Coleman's model, the evidence provided so far seems to corroborate the lower-left part of the bath-tub, i.e., the presence and relevance of links 1 and 2 in the dynamics of moral change. However, this evidence does not yet knockout pure supra-individualist views. Although the evidence I provided openly contrasts with the bold view according to which change in individual psychology plays virtually *no role* in significant moral shifts (Sauer 2019; 2023; Smyth 2017; 2020), it can still be seen as compatible with a more charitable reading of a supra-individualist view, according to which psychological moral change is possible, but only as a fortunate (and still limited) by-product of changes occurring at the level of macro socio-ecological conditions. If even just this moderate theory proved to be true, it would still make little sense to assign priority of intervention, when required, on strategies for psychological

change aimed at broader (e.g. collective, institutional) moral-social reform. My model challenges this weaker view as well. As we will see, micro-level psychological moral change is not a mere accidental by-product of supra-individual circumstances but can also be, under certain conditions, voluntarily promoted and realized, as wells as cause substantial socio-institutional moral change.[106]

The thesis put forward here is fairly simple. Certain socio-ecological conditions can favor the acquisition of resources which can facilitate agency, i.e. increased ability to consider and exert control on one's decisions, choosing among courses of action that represent alternatives to what is prescribed and/or strongly incentivized by local social norms, traditions, habits, individuals' learning history, biological pressures, and so forth (the idea of 'open-ended normativity'). Christian Welzel identifies some fundamental action resources which allow for emancipative values to rise, and for people's agency to increase, which include material and social resources and, crucially, intellectual and epistemic ones, such as greater availability of information, knowledge and intellectual skills developed through formal education, but also tools and social resources allowing for their proliferation and stability (i.e. increased territorial and relational mobility, devices such as the printing press, sufficient levels of freedom of association and expression etc.) (Welzel 2013; Smyth 2020, 15-18; see also Buchanan 2020, 146, 166).[107] These resources can substantially increase the opportunities of individuals and institutions, making them know, imagine, and desire new goals, courses of actions and ways of life, leading them to exert greater control on their own lives, environments, and decisions.

Action resources are necessary for the exercise of open-ended moral reasoning and

[106] Here it is important not to misunderstand the technical concept of 'evolutionary by-product' and simple 'by-product' in its ordinary meaning of 'secondary', 'incidental' side-effect. In the former case, the concept of by-product is not opposed to something that involves intentionality, since there is no intentionality in natural selection. A trait is an evolutionary by-product if it has no clear, direct biological function, and we can identify a plausible causal sequence explaining its development as a consequence of other evolved traits. On the contrary, in the latter case, understanding psychological and moral change as by-products of broader, supra-individual dynamics means to conceive them as unintentional outcomes, explainable without the need of referring to human moral agency, and essentially uncontrollable by acting on the micro level of individual psychology. Here, I am using the term 'byproduct' in the latter sense. It is important not to misunderstand these two possible meanings of the term also because the core theses put forward here is that a) human moral cognition and capacity for open-ended normativity can be understood as *evolutionary* by-products, while b) not any specific instance of moral change can be understood as a mere by-product of unintentional supra-individual dynamics. The objects of analysis are also distinct: one the one hand, what I see as evolutionary byproducts are human agency, moral cognition, and the capacity for open-ended normativity; on the other hand, what I think should *not* be understood as simple byproducts of macro-level structural-institutional dynamics are specific instances of moral change.

[107] In his important discussion of the role played by moral pioneers in the dynamics of moral change, Buchanan (2020, 164-167) stresses the crucial role of other, non-epistemic enabling conditions of radical moral change. Above all, it is worth mentioning the importance of socio-institutional conditions that make the revolutionary action of moral pioneers a bearable cost for themselves (otherwise, if the cost were too high, it would be hard to explain their motivation to engage in revolutionary actions). Buchanan suggests that some of the greatest moral achievements of recent history have been favored by the fact that societies were complex enough to guarantee that moral pioneers – rule-breakers and first adopters of new worldviews who in institutionally less complex societies would have been ostracized and punished – could still benefit from their participation is several other cooperative schemes. As Buchanan observes, it is no coincidence that most of the early British abolitionists came from the middle or upper-middle class (166). Thanks to Allen Buchanan for pointing this out to me.

normativity, which consist in the abilities to be critical towards received opinions, norms, habits, and to consider and choose alternative courses of action that deviate from those immediately and more strongly suggested and incentivized by specific types of learning and value representation (more on this below). Open-ended reasoning processes require several and relatively costly resources,[108] but they also allow to perform very peculiar kinds of tasks and establish peculiar types of institutions. As we will see more in detail in the final sections of this work, decades of research in cognitive (neuro)science show that moral – as non-moral – decisions can be modeled as the result of (at least) two different kinds of learning and decision-making processes (Kahneman 2011; Greene 2014). Although several characterizations of dual-process frameworks have been provided in the literature, I will mostly rely here on the distinction, based on computational models of reinforcement learning, between model-free and model-based learning and decision algorithms (Bina 2022; Cushman 2013; Dolan & Dayan 2013; Greene 2017; more on this in chapter 10).

On the one hand, certain resources such as information and knowledge[109] favor model-based learning, reasoning, and decision-making, which facilitate agency and, in turn, *prospective* and *promotion* decision strategies. On the other hand, their absence and other socio-ecological conditions hinder open-ended reasoning and normativity, favoring reliance on model-free learning and decision-making, which in turn facilitates *retrospective* and *prevention* decision strategies (Welzel & Inglehart 2010, 49).

Table 2. *An evolutionary sequence of adaptive and open-ended links explaining increase in agency and open-ended normativity*

1. Availability and environmental utility of resources (objective)
Socio-environmental conditions define (objectively) available opportunities to survive/thrive.
E.g. in agrarian societies, going to university is not an opportunity to survive/thrive; professional training is not an opportunity for women in patriarchal societies. For these reasons (cf. the etiological theory of function), higher education and professional training institutions for women will not be (very) present in these societies. Their local absence would then be another factor hampering the possibility to receive a professional or academic education, hence lowering, even more, the chances of being directly or indirectly exposed to their benefits.

[108] Some of the most important ones are the availability of factual information and knowledge and the variety of social experiences, fundamental for the possibility to develop reliable causal models of one's environment and represent possible alternative courses of action.

[109] I intend here knowledge and information in a very broad sense, including knowledge of (or information about) the probability that *x* occurs after *y*, as well as knowledge of (or information about) different points of view and ways of living.

2. Value utility (subjective)

People value (subjectively and collectively) what is most helpful in light of available opportunities.

Under different social, economic, cultural or environmental conditions, some resources and behaviors are more valued and pursued than others because they are more advantageous for surviving and thriving, and hence become more easily available (less costly) and normal (Bicchieri 2016). For example, in agrarian societies manual skills are more useful and hence, on average, more subjectively valued (in individual utility terms), more pursued, and widespread than intellectual ones.

Differences in the objective existential utility and availability of resources (e.g. relational mobility, information, knowledge) make people appreciate, pursue, and conceive as normal different kinds of actions and behaviors. Under certain conditions – e.g. in market-integrated contexts – tolerance and trust, openness to change, self-enhancement, self-actualization and self-expression are positively valued because they provide opportunities to thrive. Under different conditions these traits are disvalued because they are useless, and even costly and dangerous, while traits such as conformity, defense of tradition, intensive kinship and stronger relational bonds are subjectively valued because they (objectively) provide greater opportunities to survive and thrive.

3. Value representation – Pursuit strategies – Well-being link

When impersonal prosociality, knowledge and emancipative values – self-actualization, self-enhancement, self-expression – are appreciated and pursued thanks to their objective ecological utility, and easier to develop because of availability of enabling resources, we have the enabling conditions for greater agency and open-ended normativity.

Impersonal prosociality favors knowledge of and openness to new and alternative perspectives, ways of life, and norms. Knowledge and experience of different perspectives and alternative courses of actions, together with increased analytic, prospective, and imaginative thinking favor model-based reasoning and decision-making (MB). Individualist traits such as self-actualization, self-enhancement and self-expression correlate with lower levels of conformity with received (e.g. local) norms, rituals (etc.), and greater decisional autonomy.
Individualist traits and altruism are also considered as the highest source of life satisfaction, and this incentivizes them.

All these traits get selected because they are useful under specific circumstances, but then they can be used for other purposes.

One of them is engaging in more conscious and open-ended moral reasoning, allowed by model-based reasoning and decision-making, as a byproduct of the selection of greater skills in some of its more specific components (e.g. greater analytic skills, mental time-travel, self-control, etc.) which originally evolved for other purposes.
MB is based on imagination and consideration of alternative possibilities on the basis of a causal model (a mental 'map') of the environment, built on representations of outcomes, expected values, rewards and transition functions.

On the contrary, lack of resources, lack of knowledge and deferential values correlate with deontological and intuitive moral reasoning. Decisional inflexibility, static environments and systems, and lack of information favor model-free reasoning and decision-making (MF). MF works by associating value to available actions after a past history of rewards, independently of the representation of a causal model of the environment.

My model explains the emergence of greater agency and open-normativity as a byproduct of the selection of psychological traits explainable by research such as Henrich's.[110] The psychological changes we have seen – analytic thinking, greater impersonal prosociality and trust, individualism, mental time-travel, etc. – facilitate engagement in more sophisticated model-based moral reasoning and decision-making, which is necessary for moral agency, open-ended normativity, and justification (see sections 10.2 and 11.3 below, and Railton 2017). Where the development of these psychological traits is beneficial, the impact of individual agency on social-institutional changes is greatest, because they are freely sought after (although of course other less non-agential dynamics and causes concur).

Where, on the contrary, these abilities are not favored, e.g., by sociocultural contact and exchanges, markets, information and knowledge (both factual and emotional, experiential, etc.) parochialism, exclusivity, dogmatism, moral absolutism, closeness, injustice, patriarchal values (etc.) prevail. In these cases, individual decisions and behaviors are at the mercy of tradition, social norms, taboos, fears, and deference to authority, and it is harder for psychological moral change to drive socio-institutional moral change, since agency and open-ended normativity are more costly and reduced, and supra-individual, socio-ecological causal dynamics are more influential.

In the opposite case, however, certain traits can favor open-ended normativity, so that it can be stated that even individual moral change is itself less influenced by the social context. Certainly, the enabling conditions – the development of specific psychological traits – are more context dependent, but once individuals develop them, they are more free, i.e., more able to exert control on their lives: they can imagine, represent, value, and pursue different goals and courses of action.

Finally, even macro-level social change is more significantly driven by individual action. Under these circumstances the bottom part of the bathtub model counts more, explanatorily speaking:

> As an evolutionary process, value change involves humans as agents who make distinct strategy choices. Agency, understood as the capacity to make purposeful choices, and the learning potential connected to agency, catapult the evolutionary pace of human societies on a new level, accelerating cultural evolution way beyond biological evolution (Welzel &

[110] Again, there is no reason to believe that moral agency and open-ended normativity depend on distinctively Western cognitive abilities. It's true that these cognitive tendencies have been culturally prominent and highly valued in the West (thanks in part to prominent technology and market economies), but they exist in all cultures and have been highly prized and developed in many cultures (see footnote 84 above). Note that some kinds of MB thinking (for navigation, and perhaps other things) also exist in other species, such as rodents.

Inglehart 2010, 46; see also Ayala 2010, 9019; Singer 1981/2011).

This model challenges Buchanan & Powell's idea that open-ended normativity and agency-driven 'inclusivist anomalies' cannot be explained in naturalistic, evolutionary terms (2015; 2018).

Part III. Improving moral capacities

Introduction: Wide and narrow moral progress, again

In Part II, we saw that robust psychological change is possible, and that it can be both the driver and the result of broader, structural-institutional change. In this third part, we will focus on the idea of an improvement in moral decision-making abilities. The debate on this issue is somewhat fragmented in the contemporary literature: a great hybrid philosophical-psychological discussion is taking place in virtue theory over virtue development (Annas et al. 2016; Miller 2017; Stichter 2018; Wright et al. 2020) and practical wisdom (De Caro & Vaccarezza 2020), though this represents a very specific philosophical approach to the issue. Needless to say, experimental research in moral psychology is booming (Greene 2015) but, apart from a few exceptions – and understandably – experimental psychologists avoid messing with normative-evaluative issues and philosophical discussions.

Finally, discussion over these issues is growing in the contemporary debate on moral progress. As already stressed in Part I, however, within this literature moral progress has been so far mostly understood in a 'wide' sense, as "*any kind* of morally desirable social change" (Sauer et al. 2021), even when change is not actively driven by "the exercise of the moral powers or improvements of them" (Buchanan & Powell 2018, 51). Typically, these wide conceptions also maintain that the idea of moral progress "applies only to events, institutions, and practices in countries, cultures, societies, eras, or periods in history – not to individual persons or personal moral behavior" (Macklin 1977, 370). As I suggested in Part I, there are

good reasons to hold even instances of change that are not significantly agency-driven as instances of moral progress (though they should contribute to increasing agency as a result).

However, radically 'wide' conceptions of moral progress (Sauer 2023) conceive of moral progress in *supra-individual* and *value-externalist* terms. This means, first, that they consider social structures and institutions – rather than persons – to be the main subjects of moral improvement; second, that they locate a considerable part of the specifically moral, progressive value of social change in supra-individual structural relations or processes, or in the outcomes that they produce (e.g. increase in well-being or equality), rather than individual minds, or in the improvement of the moral capacities of individuals (Sauer 2023, Introduction).

As already discussed, structural relations between people and/or the actual production of desirable outcomes – e.g. lower aggregate levels of violence or suffering, increased population welfare, fairer formal institutions, etc. – are important aspects to be considered in diachronic moral evaluations. However, as I argued in Part I, improvements in people's agency and abilities for open-ended reasoning are particularly important aspects to be included in this kind of moral evaluation (see Buchanan & Powell 2018; Severini 2021; Songhorian et al. 2022), since improvement in these capacities as drivers is not only an important enhancer of the moral worth of progressive moral shifts, but it can also drive further emancipatory processes, stabilize them, and avoid regressive turns (see section 3.2 and chapter 7 above; see also Kumar & Campbell 2022; Welzel 2013).

As we have seen, moral agency and open-ended moral reasoning and decision-making are complex phenomena, which require favorable socio-environmental enabling conditions as well as the development and exercise of peculiar cognitive capacities. In chapters 6 and 7 I focused my discussion on the enabling conditions for psychological change that can allow greater agency and open-ended reasoning and behavior. In this part, I will narrow the focus, concentrating on even more 'micro' aspects of this issue, such as the development of psychological traits, reasoning and decision-making abilities that can contribute to enhancing individual agency.

With this, I do not mean to reduce the complexity of moral change to individual-psychologistic dynamics[111]. Structural-institutional and individual-psychological moral change interact through complex feedback loops, reciprocally influencing one another (Buchanan & Powell 2018, 210-11; Kumar, & Campbell 2022, 135-138) and, as for other social phenomena, reducing the complexity of moral change to one or the other level is problematic (Archer 1995).

[111] For a critique of methodological structuralism/anti-individualism, see Madva (2016).

Hence, if the two levels are equally important – from axiological, methodological, ontological, and explanatory points of view – I believe that research into the individual dimension of moral progress deserves at least more space than what has been dedicated to it in the contemporary debate on these issues. An account of 'individual moral improvement' seems therefore important within a theory of moral progress, and especially within an agency-based one.

As in Part I, in the following chapters I do not aim at offering a definitive answer to such a complex and millenary philosophical and existential topic. What I will more modestly show is that insights from virtue-based theories, recent research on character education and advances in moral psychology and computational models of learning and decision-making can offer fruitful insights to reflect on this issue.

Before starting, a final methodological caveat is in order. The ideas of moral progress and improvement raise a number of questions on the nature of moral facts and knowledge. Whether it is possible to legitimately speak of these issues without relying on moral realism is a matter of debate (I will briefly discuss this problem in chapter 10). In any case, I believe that my claims on the importance of individual moral progress, on the role played by learning processes, and by the development of peculiar character traits in it hold irrespective of different metaethical accounts (see Schaefer & Savulescu 2019; Songhorian et al. 2022 for a similar claim).

Part III develops as follows. In section 8.1 ('A framework for individual moral improvement') I propose three basic requirements that both theoretical accounts and concrete strategies for individual moral improvement should satisfy. In 8.2 ('Contemporary accounts of individual moral progress') I evaluate some recent accounts and strategies for individual moral progress, and show that they fail to meet these requirements. In chapter 9 ('A virtue-based approach') I discuss a two-step virtue-based account that sees individual moral improvement as relying on important aspects of moral development (9.1), but also – and crucially – as being able to *transcending* them (9.2). I then conclude by showing that this account successfully meets the proposed criteria (9.3).

Nonetheless, a virtue-based account alone does not seem to provide a sufficiently clear normative criterion to understand in which direction individual moral progress should be headed, nor how to address specific, complex and controversial moral problems. In chapter 10, I show how this conclusion is supported by recent psychological research that highlights the limits of affective learning and the reliability of moral intuitions even if 'educated' or 'trained'. In 10.2, I suggest that our increased scientific understanding of the cognitive processes involved in moral decisions can have relevant normative implications, though this does

necessarily lead us to recognize the superiority of a specific normative ethical theory: these normative implications can be conceived as procedural, rather than substantive. Finally, in chapter 11, I discuss some implications of these conclusions for the problem of assessing improvements and/or different levels of reliability in moral decision-making. I conclude, again, that even for this kind of operation a procedural approach is preferable.

8. Individual moral improvement[112]

1. A framework for individual moral improvement

As in Part I, I will start and frame this inquiry on the idea of individual moral improvement with some methodological considerations about how I believe such an inquiry should be conducted. To provide an assessment of some of the main proposals that have touched the issue of individual moral improvement in the contemporary debate, I suggest that any reasonable account should satisfy the following minimal conditions. This list of conditions is certainly not exhaustive, but I believe that it has the advantage of focusing on criteria that are relatively non-controversial and shareable by different perspectives. I will focus on three main requirements: feasibility, effectiveness and stability.

Feasibility. A relatively non-controversial methodological starting point is that moral theories should be consistent with our best scientific knowledge, e.g. with data from the cognitive, behavioral, and social sciences. This minimal naturalistic concern suggests that – to be useful both as evaluative and thought- and action-guiding tools – accounts and strategies for individual moral progress should be *feasible* (or realistic: see e.g. Flanagan 1991), i.e. both their descriptive and prescriptive elements should be consistent with our best empirical

[112] This section and the following have been re-adapted from a paper co-authored with M. Liberti, M. Reichlin, S. Songhorian, and M. S. Vaccarezza. I really thank them for their contribution. See Bina et al. (manuscript).

knowledge about humans' material, environmental, cultural, and psycho-biological constraints,[113] and should be practically implementable by moral agents.

One of the main reasons why the recent debate focused on broader, socio-institutional aspects of moral progress can be found in the rise of 'naturalized' approaches to the study of moral cognition and behavior. Recent experimental research has contributed to emphasizing the influence of constraints and biases on human computational, deliberative, perceptual, affective, and motivational capacities (Dickert et al. 2012; Doris 2014; Haidt 2012; Kahneman 2011; Klenk & Sauer 2021; Schwitzgebel & Cushman 2015), while evidence from history and evolutionary anthropology shows that several cultural and socio-institutional shifts that we tend to evaluate positively (at least in WEIRD societies) occurred without being consciously designed or agency-driven (e.g. Acemoglu & Robinson 2012; Fukuyama 2012; Henrich 2015; 2021; North et al. 2009; Rehren & Sauer 2021).

As I have in the previous sections of this work, I share such a naturalistic approach, believing that moral theory should be significantly informed by scientific knowledge and evidence, though not entirely reduced to it (see Part 1; De Caro & Macarthur 2010). Nonetheless, as I will argue, I do not believe that committing to such a naturalistic methodology casts doubt on the possibility of robust individual moral improvement, as some critics have suggested (Doris 2014; Klenk & Sauer 2021; Sauer 2019; 2023).

Effectiveness. Second, if we want accounts and strategies for individual moral progress to be useful *prospectively* and not merely *retrospectively* (Bina 2022; Greene 2017; Railton 2017) – i.e., tools to identify and promote further moral improvements, rather than just to evaluate already occurred shifts – these projects should be effectively thought- and action-guiding. In other words, they should point out effective strategies for the promotion, realization, and consolidation of improvements in individual moral capacities, agency and decisional autonomy, such as better moral reasoning and motivation, compliance with valid norms, conceptual competence, sympathetic imagination and understanding (see chapters 10 and 11 below for greater discussion of these capacities; see also Buchanan & Powell 2018; Schaefer & Savulescu 2019; Songhorian et al. 2022).

In line with my 'dual' agency-based account of moral progress outlined in Part I (chapter 3), this minimal requirement of effectiveness should apply to 'internal' changes (in beliefs, reasoning, and sensitivity) as well as to 'external' ones (actual behavior change, outcomes): a

[113] These constraints, however, should not be conceived of as too rigidly fixed. See Part II of this work and Buchanan & Powell (2015).

complete account of moral progress should be able to consider both, and good strategies to foster progressive individual moral change should be able to effectively produce both. This integrationist approach can lead to cast doubt (a) on hyper-intellectualist accounts that insist, mostly or exclusively, on the improvement of epistemic abilities which do not necessarily translate into behavior (see also Part I; Persson & Savulescu 2017, 289-290)[114]; and (b) on accounts treating unintended or fortuitous changes in attitudes, feelings or behavior (e.g. driven by unconscious mechanisms or non-moral motivations) as instances of robust individual moral improvement (more on this below).

Stability. Accounts and strategies for individual moral improvement should aim to identify and promote moral changes that are sufficiently stable, i.e. not episodic. Arguably, nobody would consider entirely occasional, contextual, or accidental behavior change as instances of moral improvement (Schinkel & de Ruyter 2017, 134). Improvements in moral capacities should be consistent across contexts and domains, they should last in time and, if possible, keep progressing.

It is important to notice that, as emphasized by Rehren & Sauer (2022), stability can reinforce moral progress, but it can also hinder it. Indeed, stability should not be understood in terms of rigidity or inflexibility, neither in a developmental-psychological nor in a substantive normative sense. As I argued in Part I, and as I will stress below, to adjust judgments and behaviors to relevant information and contextual changes, or being disposed to revise one's moral and non-moral beliefs are necessary conditions for individual moral progress (Greene 2017; Schaefer & Savulescu 2019, 78). Such open-ended traits, that John Dewey called "habits of flexibility" (Johnson 2022), can be stabilized and even enhanced, although they precisely consist in being open to alternatives, and to revising one's beliefs and behavior.

2. Contemporary accounts of individual moral progress

In the previous section, I suggested that feasibility, effectiveness and stability are three basic and relatively non-controversial requirements that accounts and strategies for individual moral improvement should meet. In what follows, I evaluate some influential contemporary accounts and strategies for moral improvement and show that they all fail to meet at least one of these criteria.

[114] As also Kumar and Campbell notice, "improvements in moral emotions, norms, and reasons must be effective. No moral progress worth its salt occurs if people's ideas are ennobled, but nothing on the ground changes" (2022, 179).

I start by looking at some contemporary theories of moral progress, dividing them into theories that understand the individual dimension of moral progress as improvements in epistemic abilities (e.g. changes in moral beliefs, reasoning, and understanding), and theories that understand moral progress as changes in attitudes, sentiments and emotional dispositions triggered by several social and emotional experiences (e.g. traumatic, aesthetic). I then look at more concrete, empirically-informed strategies targeting individual psychology with the aim of steering people's motivation and behavior in a better direction. I conclude that all these accounts and strategies fail to meet the aforementioned desiderata.

As I discussed in Part I, 'narrow' accounts of moral progress consider change as morally progressive only when it is considerably agency-driven, i.e. when it involves the active exercise or improvements in beliefs, reasoning, and/or deepened understanding of moral problems and concepts. For instance, Moody-Adams (1999; 2017) identifies improvements in the understanding of the semantic depth of moral concepts as one of the main drivers of moral progress (cf. Platts 1988). In particular, she distinguishes changes in beliefs from changes in practices, and argues that it is the former that mostly drives the latter: it is when individuals come "to appreciate more fully the richness and the range of application of a particular moral concept (or a linked set of concepts)" (1999, 169) that beliefs change for the better; and it is when such understanding is "concretely realized in individual behavior or social institutions" (ibid.) that practices change accordingly.

This focus on individuals' ability to deepen their semantic understanding of moral concepts (e.g. justice), however, has been accused of being reductionist and too intellectualist (Buchanan & Powell 2018, 61)[115], individualist, and unrealistic (see Hermann 2019; Klenk & Sauer 2021). For current purposes, it suffices to note that if moral progress in practices comes after some individuals have gained better grasping of moral concepts, then we would need to shed more light on how such epistemic success is obtained, as well as on how epistemic improvements can relate to actual behavior change.

Peter Singer's view (1981/2011) can be seen as another narrow account of moral progress. According to Singer, the extension of moral concern beyond one's kins and tribe, and the possibility of revising morally problematic cultural norms, can be driven only by reason (Singer 2011, chapters 4-5), which is also the only way to acknowledge the validity of impartial and universalist moral principles. Singer dismisses the idea that change in sentiments can lead to a

[115] According to Buchanan & Powell (2018), mere improvements in moral concepts and reasoning are "instances of epistemic progress but are moral progress only insofar as they contribute either to better compliance with moral norms or to better motivations or the flourishing of the moral virtues" (92).

comparable level of moral inclusion and consideration. However, while mostly discussing why certain moral principles should be seen as objectively more rational than others, he does not specify how concretely people can come to know, appreciate, endorse, and justify them; how agents can develop motivations that facilitate their practical implementation; how these traits can be cultivated and made effective; nor how people can develop skills enabling them to perceive, reason, and decide more reliably in specific contexts.[116] Singer's view is problematic not only for normative and metaethical reasons, but also because it appears practically hard to implement. Like other 'narrow' epistemic accounts, Singer seems to assume a controversial position about motivational internalism, failing to acknowledge that improvements in the impartial consideration of interests do not necessarily translate into actual feeling, motivation, and behavior.

Being about moral progress – and not specifically *individual* moral progress – the accounts just discussed also consider moral change and progress in their broader, aggregate, societal, structural dimensions. But their emphasis on psychological, epistemic and deliberative aspects of moral change makes them plausible candidates for being considered accounts of individual moral improvement as well. However, to summarize, these accounts suffer from two main problems: first, none of them specifies how shifts in reasoning, beliefs or understanding – e.g. more open-ended, internally and externally consistent – can actually occur, for instance by referring to available empirical research. Belief change and deepened understanding of moral concepts are certainly possible, but these views provide almost no empirical evidence as to how and to what extent improvements in moral reasoning and decision-making abilities can actually take place (for an empirically-informed skeptical view, see Klenk & Sauer 2021; see also Haidt 2001; 2012; Schwitzgebel & Cushman 2012; 2015). Hence, these theories do not meet the *feasibility* desideratum. Second, these accounts also fail to meet the *effectiveness* requirement. Even accepting a mildly optimistic view about motivational internalism, intellectual and epistemic improvements on moral issues do not always correspond to/produce changes in actual behavior.[117] Excessively rationalist or intellectualist accounts seem to miss something important in their analysis of the dynamics driving individual moral improvement. What about sentimentalist ones?

Although not always explicitly framing the issue in these terms, some scholars have provided alternative pictures of the dynamics of individual moral progress. Grounding their

[116] Except for a fairly acritical endorsement of Kohlberg's account of moral development, and the (over)confidence that "there is a demonstrable connection between moral reasoning and moral action" (Singer 2011, 138).
[117] A mismatch highlighted, for instance, by recent experimental results on the moral behavior of ethicists (Schwitzgebel & Rust 2016).

views on quasi-Humean accounts of moral psychology, they claim that changes in moral judgment and behavior can only be triggered by affective-emotional experiences (Rorty 1999; Slote 2007). For instance, Rorty argued for a *progress in sentiments* (129) rather than in epistemic and intellectual capacities when it comes to changes for the better in human relations. Although placed at the opposite end of the spectrum if compared to intellectualist accounts, these views also risk failing to meet the stability desideratum. In fact, the kind of psychological moral shifts conceived by these views depend primarily on contingent affective experiences, rather than on considered reasons and beliefs, deeper and more systematic understanding of morality, or the development of stable dispositions to act.[118] Moreover, as it was the case with intellectualist views, it is not clear which concrete strategies should be adopted to effectively allow such improvements since, within these 'emotivist' frameworks, individual moral change appears hardly under conscious control.

Psychological research has recently offered interesting insight into this discussion. In addition to philosophical accounts, in fact, more concrete and empirically-informed strategies have been developed to foster progressive change by operating, firstly, on the individual-psychological level of beliefs, empathy, emotions, and motivation. I will focus here on moral bio-enhancement (MBE) and nudging as two very influential strategies for improving individual behavior.

MBE has been widely discussed in recent years as an empirically-informed and promising strategy to improve human moral capacities such as empathy, benevolence, fairness, or motivation by means of pharmacological, genetic, and/or other bio-technological interventions (Crutchfield 2021; Douglas 2008; Persson & Savulescu 2012). There are good reasons, however, to doubt that MBE meets both the effectiveness and stability desiderata. Even if, in fact, biomedical interventions were actually effective enhancers of morally relevant capacities, – such as empathy, trust, fairness, cooperation –, it is far from granted that such changes would constitute genuine instances of individual moral progress: biomedical enhancements may increase the probability of certain behaviors and outcomes deemed desirable, but this strategy risks bypassing (and even weakening) individuals' capacities to choose autonomously based on what they consider the most convincing moral reasons, rather than being pre-programmed

[118] A middle way between these two extremes has been suggested by Severini (2021). Severini defends a debatable first-personal, Humean sentimentalist view that attributes no *ex-ante* causal role to reasoning and reflection in the process of deepening our moral understanding (that is, for her, what moral progress consists in): "on my account what can improve one's moral understanding of a certain situation consists of being actually (or imaginatively) involved in that situation firsthand. So, for instance, after having witnessed animal suffering in a slaughterhouse, we gained a better understanding of why killing animals is wrong. It is the experience of the unpleasantness of the animal suffering which leads you to believe that killing animal is wrong and not the other way round, i.e. it is not the case that the general claim that killing animal is wrong makes you aware that the animal suffering you witnessed is wrong" (101, footnote 58).

to judge and act according to specific normative standards (Harris 2012; Reichlin 2019). MBE interventions fail to meet both the effectiveness and the stability requirement, since they do not allow improvements in the autonomous, critical exercise of moral agency.

A second empirically-informed project aiming at fostering better decisions and behavior is nudging. Nudging is a technique conceived to promote good values (Thaler & Sunstein 2008) by triggering people's automatic and often unconscious cognitive processes (Kahneman 2011). Although this approach was not originally developed nor framed as a strategy for individual moral progress, nudges are also designed to push people, more or less gently and transparently, to change their behaviors, habits and beliefs towards other people, society, or the environment, with a clear ethical purpose (Sunstein 2015).

Like MBE, nudging mostly bypasses people's agency and decisional abilities, rather than improving them. Moreover, being less invasive and more contextual than MBE – nudging works by changing choice architectures and physical affordances, rather than acting on neurotransmitters, hormones or peptides –, it fails to meet the stability desideratum. In fact, there is no guarantee that people's behavior and judgment will be consistent in the absence of the stimulus (or in front of different choice architectures). Finally, nudging fails to meet the effectiveness requirement. On the one hand, continued exposure to nudges can produce habituation effects; on the other, the more transparent and informative they are, the lower their impact in producing the desired behavior (Casal et al. 2019).

9. A Virtue-based approach[119]

So far, I have shown that much can still be done to offer a robust, empirically-informed account of individual moral improvement, which is needed both in itself and in relation to societal moral progress. In this section I will show that, although they cannot be considered the *only* possible tool to address this issue, virtue-based theories can and should have a seat at the table of this debate. Above all, a virtue-based account of individual improvement focuses on the necessity of developing peculiar psychological traits for enhancing agency and decisional autonomy, while also providing reliable criteria to regulate their exercise. In this way, a virtue-based account is able to respond – at least in part[120] – to the critique of arbitrariness of an agency-based account of moral progress considered at the end of section 3.2.

In what follows, will argue that a virtue-based perspective can successfully satisfy the aforementioned requirements, by providing a robust account of individual moral improvement which allows for increased agency and decisional autonomy (in this way, such a virtue-based account is perfectly consistent with the agency-based account proposed in Part I). This account can be intended as a two-stage process. The first stage emphasizes the importance of moral development, which I take as a necessary – yet insufficient – condition for moral improvement. The second stage requires that the substantive components of the moral outlook of a mature agent could change, even dramatically, in order to attain moral improvements.

[119] I am particularly grateful to Maria Silvia Vaccarezza for her considerable contribution to this part.
[120] As we will see in chapter 11, in fact, to fully respond to that critique we need also to refer to certain procedural constraints.

Thus, I claim that a virtue-based account of moral improvement can provide: i) A robust account of moral development; ii) The categories to explain how moral development can be transcended, and one's moral and intellectual character re-shaped taking into account new experiences, challenges, and information. As to i), neo-Aristotelian approaches to character education can provide both the theoretical presuppositions and an empirically-minded account of the main stages of moral development. For what concerns ii), the issue is relatively underexplored. This contribution aims precisely at addressing this need, by highlighting how virtue theories can offer a flexible, open-ended conception of individual moral improvement. In section 9.2 I will argue that virtue-ethical conceptual tools can offer fruitful insights for this operation. Finally, in section 9.3 I will show how the account proposed meets the criteria discussed above, and can therefore count as a satisfying theory of individual moral improvement.

1. Virtue-based moral development

Until a couple of decades ago, Neo-Aristotelian character education was an empty label, generically opposed to Kohlbergian models of moral development. However, largely due to David Carr's pioneering work (Steutel & Carr 1999), and to the subsequent efforts of many neo-Aristotelian philosophers, nowadays the theoretical and empirical landscape has dramatically changed. To the point that it is now possible to flesh out both the basics of an independent virtue-based account of moral development – corroborated by empirical evidence and widely tested in projects developed in schools – and the stages of moral development in terms of virtue acquisition.

The main theoretical presuppositions underlying a virtue-based account of moral development can be summarized as follows (see Jubilee Centre for Character and Virtues 2022):

(i) Flourishing is the aim of character education. To flourish is not only to be happy, but to fulfill one's potential;

(ii) Flourishing requires acquiring intellectual, moral, and civic virtues;

(iii) Therefore, acquisition of the virtues or their constituents is the specific aim of character education – i.e., priority is assigned to aretaic notions over deontic ones (Steutel & Carr 1999, 7). Thus, education is not primarily concerned with grasping *the* right principles of conduct and applying them to the situation at hand, but with the "cultivation of a range of sensibilities to

the particularities of moral engagement, involving crucial interplay between the cognitive and the affective" (Steutel & Carr 1999, 12).

(iv) No definite list of virtues can be drawn; these lists can vary depending on social circumstances or developmental stages. However, following the framework elaborated by the Jubilee Centre for Character and Virtues, we can agree on four main categories of virtues that are needed to promote individual and social flourishing: intellectual, moral (e.g. courage, honesty, compassion, humility, respect), civic, and performance virtues.

(v) Virtue development has at least three main components: virtue-reasoning (or knowledge component), virtue-emotion, and virtue-perception;

(vi) Finally, *phronesis* or 'practical wisdom' has an integrative function over all the virtues. It enables to "perceive, know, desire and act with good sense" (Jubilee Center for Character and Virtues, 2022).

These main tenets suggest that moral development can be conceived of as a preliminary, yet necessary step towards individual moral progress conceived as an improvement in agency. It comprises both reflective and affective elements, it requires time, exercise and habituation, but it also allows for flexibility in identifying new virtues or new domains of their application (therefore avoiding conservatism and rigidity), and it culminates in an autonomous ability to perceive the morally relevant features of situations, and to act according to one's reasons. Also, the framework suggests that developing specifically *moral* character traits is only part of a story of personal development, which also comprises the development of other traits – above all, cognitive and intellectual ones.

However, to propose a credible path towards moral improvement, this framework must be able to flesh out specific steps towards moral maturity. In this respect, a classic description of such steps is provided by what we may call a classic 'neo-Aristotelian' account, which articulates moral development into the three fundamental steps of 'natural character', 'developmental character', and 'second nature' while remaining close to Aristotle's labeling.

This account, paradigmatically fleshed out by Kristjánsson (2007; 2015) and Sanderse (2015), recalls the classic Aristotelian steps and identifies four stages of moral development: moral indifference, lack of self-control (*akrasia*), self-control (*enkrateia*), and proper virtue (Sanderse 2015, 386). While moral indifferents hold an 'amoral' view of happiness and the good (NE 1095 a22-23), those who lack self-control have a grasp of the virtues only on a 'theoretical' level: lack of habituation, insufficiently educated perception and emotions, and lack of self-assessing emotions impede them from 'going practical', i.e., internalizing virtues, judgments, norms, and principles and acting accordingly. The self-controlled have overcome

akrasia, but are, so to speak, on the 'threshold' of virtue. They experience conflicting emotions, and they struggle against inclinations contrary to virtue; they are 'almost there', but they still find virtuous acts difficult and unpleasant. Finally, virtuous agents have successfully built sufficiently stable character traits.

As anticipated, what enables one to climb the developmental ladder is the pursuit of two main pathways. The first, emphasized by Aristotle himself (NE I, 7, 1098b4), is habituation, a process of repetition of virtuous acts which requires practice and time, not unlike the process of learning to play an instrument. The second, more recently deepened by character educationalists, is role-modeling. According to Kristjánsson, "[c]haracter is caught through role-modeling and emotional contagion" much more than being learnt by studying lists of abstract values (Kristjánsson 2015, 21; see also Buchanan 2020, 154 on the role of exemplars in the dynamics of moral change).

Notice that, so far, this account provides no 'deep' normative-evaluative criterion for moral improvement: it is, rather, part of the 'realistic descriptive theory of moral change' that in chapter 1 I argued any theory of moral progress needs to specify. In this sense, a virtue theory is fundamental within an agency-based theory of moral progress, since it can offer a set of fundamental tools to understand and promote the exercise of greater agency and decisional autonomy.

Without parting from the general structure outlined so far, other accounts have been proposed to the effect of translating Aristotelian intuitions on moral development into more up-to-date psychological language. Among these attempts, it is worth mentioning the so-called 'skill model', which makes use of the language of skill acquisition to conceptualize the mechanism of virtue development. To summarize, Annas (2011), Hacker-Wright (2015), and Stichter (2007; 2018) agree on a framework that considers the acquisition of virtue as a process that requires practice and repetition up to the achievement of a certain degree of automaticity that signals that its owner has become skillful. Thus, acquiring a virtue is analogous to becoming able to play a musical instrument, drive a vehicle, or cook well. This can be achieved via a *training stage* of mechanical repetition of gestures, which must be internalized and assimilated, and an *expertise stage*, where one can act autonomously and almost automatically without any further need for direct instruction, nor for explicit reflection (see also Sauer 2012).[121]

Both the classic neo-Aristotelian and the skill model of virtue acquisition point to a

[121] Drawing on Bina (2022) and Greene (2017), in Chapter 11 I will discuss a significant limitation of this model.

development which promises to be feasible, effective and stable[122]. However, as mentioned at the very beginning of this work and widely discussed in Part I, moral development and character education do not *necessarily* increase agency and decisional autonomy. Some of these processes might lead to building character traits which constrain people's agency and opportunities (both of their bearers and of others), and/or consolidate values that belong to a problematic *status quo* (see also Rehren & Sauer 2022). Not any kind of character education is agency-enhancing: we need a sufficiently determined criterion to avoid the development of such constraining traits, but also sufficiently flexible to allow for open-endedness. In the next section, I will suggest that contemporary virtue theories are able to account for the existence and promotion of traits that can allow for such a demanding and delicate balance between stability, effectiveness, efficiency, and flexibility.

2. Transformative virtues

As I stressed in the first chapter of this work, while moral improvement and moral development are closely related – and many of the 'normal' transitions in moral development can be judged positively – moral development and moral improvement are different concepts. As stated, the concept of moral development descriptively refers to a process that can also take directions we may reflectively evaluate as negative, or problematic, according to some given normative-evaluative standard; on the contrary, moral improvement, like progress, is an inherently positive evaluative concept. Moreover, flattening the idea of moral improvement even on a 'positive' model of moral development would be either too vague and limited in its ability to provide helpful guidelines to cope with moral problems, or would make it collapse into a 'determined fixed account' (cf. Buchanan & Powell 2018), unable to account for the possibility of revising problematic aspects of one's development that could emerge in the future. Hence, a virtue-based account of moral improvement towards greater agency and autonomy cannot do without a further step, i.e., accounting for the possibility of transcending one's own development and learned values, principles, beliefs, motives, habits, and so forth (cf. Habermas 1990).

First of all, there is widespread agreement among neo-Aristotelians that a satisfactory account of human flourishing must be open to changes and revisions (Swanton 2016, 125;

[122] Obviously, this is not the only empirical conceptualization of virtue acquisition; for a recent alternative, see e.g. Wright et al. (2021).

Kristjansson 2020; Snow et al. 2021), made possible both by autonomous self-reflection, by exposure to different moral perspectives (cf. Campbell & Kumar 2012; Campbell 2017; Habermas 1990; Kitcher 2011; Sauer 2017) and by epiphanic or transformative experiences, often associated with the encounter with morally exemplary individuals (Chappell 2022). As many have noted, especially after Zagzebski's seminal work (2017), witnessing the moral excellence of a person often triggers exemplarity-related emotions capable of activating self-transforming action tendencies. Such emotions, well-known by psychologists as well as by philosophers, can be positive, such as admiration, elevation, awe (Algoe & Haidt 2009; Kristjánsson 2017; Zagzebski 2017), but also negative, paradigmatically envy, shame, and guilt (Vaccarezza & Niccoli 2018). This composite set of emotions, when activated by exposure to exemplars, shares common action tendencies, mainly emulation and inspiration. It is important to note that it is peculiar of moral exemplars to incarnate more than an outstanding level of moral development according to existing standards (even if this is sometimes the case); in their most significant instances, they rather afford new ways of enacting old virtues or lead the way to conceptualizing new ones (Colby & Damon 1992; Zagzebski 2017, chapter 8).

However, epiphanic experiences triggering self-transformation need not be necessarily related to encounters with morally outstanding humans. Several other experiences can be epiphanic or 'peak': think about transformative choices bringing about irreversible shifts in one's life, such as choosing to become a parent or to quit one's job to embrace a new life; strong unintended experiences that shake one's life plan dramatically, such as the sudden loss of a dear one, witnessing the devastating effects of a war, or – on a more positive note – falling deeply in love with someone. Or think of less dramatic encounters or discoveries that strongly brings individuals to revise their judgments, behavior and – with time – even emotions, as it has been happening over the past decades to many homophobes who discovered that someone they loved or really care about is gay (Campbell & Kumar 2022, 212-213; Kumar & Campbell 2012). Think, lastly, of encounters with 'counter-exemplars', i.e., morally despicable individuals, whose moral degradation strikes as a warning. All these experiences, as diverse as they may be, bear the potential of putting one's set of values and priorities into question, and therefore, such as in the case of encountering positive exemplars, call for a revision of one's conception of the good, right, and what a flourishing life is.

In addition to these common ways of achieving moral change, however, a virtue-based account of moral improvement would be unsatisfactory unless it could make sense of how a virtuous character has its own internal resources to allow such shifts as a result of self-reflection. In other words: if a virtuous character was a static, self-conservative psychological

set-up, paradigm shifts would necessarily be *imposed on* character, rather than *developed out of* it. Also, it would not be entirely clear what psychological mechanisms encounters with examples and epiphanic experiences would allow one to reform one's character; as well as which criteria one could use to assess when and how to address inconsistencies, conflicts, and/or new unclear situations.

A whole set of virtues identified by virtue ethicists and virtue epistemologists point precisely to the direction of incorporating – so to speak – into character a potential to disrupt existing paradigms and to revise them. These 'transformative virtues' of openness and flexibility are specifically related to practical reasoning, as well as to balancing one's existing reasons for action with the requirements posed by new situations, that can call for a revision of those reasons and our more general representation of value.

According to (certain) virtue ethicists, 'practical wisdom' is the central meta-virtue that performs functions which are critical to the paradigm shifts just outlined (De Caro & Vaccarezza 2021; Kristjánsson et al. 2021; Kristjánsson & Fowers 2022). In fact, all models agree on ascribing to practical wisdom several crucial functions, which include: the ability to perceive the salient ethical features of a given situation; the role of adjudicating among conflicting moral requirements of given situations – and, more controversially, among conflicting goals of different virtues; an orientation to the goal(s) constituting a good and flourishing life; some form of emotional regulation, which brings emotions in line with one's moral judgment (see Darnell et al. 2019, 18-20; De Caro et al. 2021; Snow et al. 2021).

A useful example of how these functions work for our present purposes is precisely that of an agent undergoing an epiphanic experience, which can reveal to the agent new ethical affordances, new ways of dealing with situations, and so on. A cognitive dissonance is produced, and existing values, behavioral schemas, principles, and idea of a good life and its components are shaken and questioned (see Campbell & Kumar 2012; Kurth 2016; Moody-Adams 1999).

However, this is only part of the story, and further conceptual resources are provided by virtue epistemology. One could object to this move that intellectual virtues are specifically directed to epistemic ends and goods, and/or that they can just work as tools to select better means, not ends, or to endure better performance; therefore, they cannot be part of the transformative and open-ended picture that I am drawing. However, while intellectual virtues alone might not be enough for moral improvement (see Buchanan & Powell 2018, 92; Persson & Savulescu 2017), moral virtues alone might not be enough either: both elements seem to be necessary. Epistemic and intellectual virtues still remain of the utmost importance, given that

re-shaping one's value representation is a hard epistemic challenge, which has to do with assessing, above all, facts, reasons, and inconsistencies.[123]

In an influential taxonomy of intellectual virtues – specifically construed to meet educational needs – Baehr (2011, 21; revised in 2021) has identified six sets of intellectual virtues, grouped according to the specific challenge to moral inquiry they enable to meet successfully. Two of these sets are particularly relevant for the current discussion. The first includes virtues of intellectual 'wholeness' such as intellectual integrity, honesty, humility, transparency, self-awareness, and self-scrutiny, which all play a key role in questioning one's beliefs and values in light of new evidence (2011, 19).

The second set includes virtues of 'mental flexibility', such as imaginativeness, creativity, intellectual flexibility, open-mindedness, agility, and adaptability. These virtues are required when one "confronts a subject matter that [...] is foreign to her usual way of thinking. Here what is needed is an ability to 'think outside the box'." (Ibid.). Because of the role these virtues can play in facilitating and regulating agency and decisional autonomy, virtue-based theories which attribute a central role to them seem to provide a fruitful synthesis between stability, flexibility and open-endedness which make them promising candidates for an account of individual moral improvement, as well as an important component of an agency-based theory of moral progress more broadly.

3. Evaluating virtue-based moral improvement

In the previous sections, I have proposed a two-step account of individual moral improvement which relies upon a virtue-based conceptual framework. It is now time to evaluate this account against the minimal requirements – feasibility, effectiveness, and stability – that I suggested at the beginning of this chapter.

As far as *feasibility* is concerned, a virtue-based account could be accused of empirical implausibility, therefore of violating even a minimal naturalistic requirement. This has been, since the 2000s, the spirit of the situationist challenge to the notion of character, according to which character traits are explanatorily irrelevant or do not even exist (Doris 1998; 2002; Harman 1999), given the innumerable situational conditionings we all are continuously

[123] As we will see more in detail in the last chapter, while being more costly, the learning and reasoning mechanisms underpinning more controlled and informed revisions of one's values and goals ('model-based' learning, reasoning and decision-making) are more flexible than value change via habit-formation and affective learning (typically guided by value-free mechanisms). See Cushman (2013), Greene (2017).

influenced by. However, over the last two decades, virtue ethicists have taken this challenge very seriously, and developed empirically credible accounts of virtue development in response. The skill-model of virtue mentioned above is only one among many different ways of re-conceptualizing virtue to meet this challenge (e.g. Alfano 2013; Miller 2014; 2017; Snow et al. 2021). The same holds for phronesis research, which is recently moving a step forward towards empirical maturity (see Darnell et al. 2019; 2022; De Caro et al. 2021). All in all, I believe virtue-based theories – their divergences notwithstanding – have successfully proven capable not only to meet the situationist challenge, but to express their core concepts in an empirically-informed language, which ensures the feasibility of virtue development as they conceive it.

For what concerns *stability*, this is one of the main strengths of a virtue-based account, since it lies at the core of any virtue theory, in line with the Aristotelian spirit (well summarized by the saying that "one swallow does not a summer make, nor one fine day" NE 1098a18). This is precisely because virtue acquisition requires time, and the proper mark of genuine – as opposed to 'natural' – virtue is that of granting a long-lasting, reliably virtuous behavior, not easily shaken by circumstantial variation (NE 1144b14). As Alfano (2013) puts it, stability as a defining feature of virtue implies that "if someone possesses a virtue at time t, then ceteris paribus she will possess that virtue at a later time t¢" (236). Among the other claims that constitute the hard core of a virtue-based story of moral development, one is particularly relevant to stability, i.e., consistency, according to which "if someone possesses a virtue sensitive to reason r, then ceteris paribus she will respond to r in most contexts" (ibid.).

The *effectiveness* requirement, finally, is met, first and foremost, by the predictive and explanatory power that comes with the possession of a virtue, to the effect that "if someone possesses a virtue, then reference to that virtue will sometimes enable one to explain her behavior", and "if someone possesses a virtue, then reference to that virtue will sometimes enable one to predict her behavior" (ibid.). Explanatory and predictive power have been the specific target of the situationist challenge; however, as stated, virtue-based theories have developed adequate responses to it. Therefore, feasibility and effectiveness are ensured by the very same empirical-mindedness which has become a landmark of contemporary virtue theories after situationism.

Virtue-based accounts of moral development, in sum, aim at fostering a development which promises not only to be feasible and to remain stable, but to effectively and reliably affect behavior. And the same features apply to the critical, open-ended character of the conception of moral improvement suggested here, which is brought about, precisely, when specific traits of character develop as stable and flexible at the same time, allowing improvements in agency

both as an ability to be consistent with one's principles, judgments and behavior, as well as to revise them in light of contextual changes, new information, and experiences.

10. Improvements in moral decision-making capacities

In the previous chapter I discussed a virtue-based account which provides an empirically-informed, robust, and action-guiding account for individual moral improvement. Such a view, I argued, provides a good synthesis in virtue of its being able to account for the development of capacities which allow for both stability, consistency, and motivation, as well as for the possibility to revise and deviate from one's values, habits, and principles in light of new experiences, information, and critical scrutiny. Moreover, the growing empirically-informed maturity of this approach provides more and more data and tools for building concrete strategies to develop these capacities in a consistent way. In this sense, such a view is able to allow for improvements in agency in two very relevant ways: that is, not only in terms of intellectual and analytical skills, but also in terms of their actual translation into consistent behaviors, intuitions, perception, sentiments, and emotions. Such a combination of improved intellectual abilities and emotional and behavioral regulation is a great indicator of increased agency, since it allows an increased control over one's thinking, goals, actions, desires, emotions, and actual behavior over several external influences and constraints. Several questions, nonetheless, remain open.

First: is it really possible to develop habits, skills, intuitions and sensibilities that are sufficiently reliable in most circumstances? Should we trust them as normative guides when addressing complex or new moral problems? If we allow for the possibility to develop accurate and reliable skills and intuitive responses in moral judgment and decision-making – as it happens in other, non-directly morally relevant domains – perhaps we should acknowledge the

existence of moral experts, i.e., people with educated intuitions and greater skills and competence which make them reliable moral guides in most situations – especially controversial ones.

This chapter develops as follows. In the first section (10.1, 'Models of moral decision-making') I present a very lively debate in contemporary psychology and empirically informed philosophy concerning the relative reliability of different types of cognitive processes involved in (moral) learning and decision-making, and briefly discuss a promising, updated kind of dual-process models of moral cognition. In the second section (10.2, 'Normative relevance') I consider some normative implications of this research by suggesting their procedural, rather than substantive, nature.

1. Models of moral decision-making[124]

Notably, decades of research in experimental psychology have been regarded by many as supporting dual-process theories of human cognition, according to which two types of processes – one automatic (type 1), the other controlled (type 2) – are involved in the psychology of judgment and choice (Kahneman 2011; Evans & Stanovich 2013). Dual-process frameworks, however, are controversial, both in descriptive terms and for their potential normative implications. Specifically, disagreement persists about the interactions between type 1 and type 2 processes and their relative reliability. I will refer to the problem of drawing normative conclusions from a better understanding of decision processes as the 'normative challenge'.[125]

According to dual-process views, type 1 processes provide quick and efficient solutions to ordinary problems. However, these responses are often statistically inaccurate, biased, and unreliable in front of new and complex problems and decisions, due to their inflexible dependence on limited information and insensitivity to new and/or relevant ones (Kahneman 2011)[126]. On the contrary, type 2 operations are more flexible and sensitive to new and relevant information and changes in the decisional environment; they are also responsible for hypothetical thinking, simulation of alternatives, and cost-benefit analyses (CBA). This, of course, requires higher computational costs.

[124] This section has been adapted from Bina (2022).
[125] Evans calls unjustified inferences from description to normative conclusions about reasoning 'normative fallacies' (Evans 2019, 6).
[126] At least at the time of decision. As discussed below, type 1 processes are not completely inflexible, since they can significantly learn over time; the point is that they cannot be updated in real time.

The idea that these differences render type 2 more reliable than type 1 processes has been widely criticized. In particular, critics have emphasized a greater interaction between processes, suggesting that the dual-process image is not accurate (Kruglanski 2013) and that type 1 processes can be subject to sophisticated learning mechanisms, made sensitive to relevant information, and attuned to considered normative standards. Controlled processes can in fact be translated into automatic ones both implicitly and through exercise, as it happens for skill-acquisition and expertise in several domains (Hogarth 2001; Kahneman & Klein 2009). In light of their flexibility, penetrability, and ability to learn, it has been argued that type 1 processes should be considered very reliable in guiding decisions (Gigerenzer 2007).

In what follows, I will explain why these reasons are not sufficient to consider type 1 processes reliable, especially to address new and complex problems, and specifically in the moral domain. This claim is based on a vindicatory etiological and procedural reply to the normative challenge: the reliability of decision strategies is assessed in light of new (non-normative) understanding of the basic processes underlying their functioning, combined with relevant features – e.g. novelty, uncertainty, stakes – of the problems at hand.

Dual-process models have been very influential also in recent (neuro)psychological research and empirically-informed ethical debates on moral judgment and decision-making. In the past two decades, empirical studies on (in)famous moral dilemmas have found correlations between characteristically deontological (D) responses and type 1 processes, while characteristically consequentialist (C) judgments correlate with type 2 reasoning (Conway & Gawronski 2013; Greene 2014; Patil et al. 2020).

A few scholars have concluded that these data support consequentialism as a normative theory (Greene 2014; Singer 2005). In section 2, I suggest that this conclusion is problematic. Nonetheless, I will argue that empirical research and updated dual-process frameworks can still support significant conclusions for moral theory, though the nature of these conclusions is *procedural* rather than *substantive*.

A big part of the recent scientific and philosophical debate has questioned both Greene's dual-process account and the normative implications that he drew from it. Many critics have stressed that type 1 and 2 processes interact much more than Greene acknowledges; that empirical evidence does not show strong correlations between D judgments–type 1 processes and C judgments-type 2 reasoning; and that type 1 processes can learn and be reason-sensitive, attuned, educated, or trained. For these reasons, critics conclude, type 1 processes are more reliable than Greene maintains (Cecchini 2021; Sauer 2017; Railton 2014, 2017).

Although these claims are true from a descriptive point of view, inferring from them that

type 1 processes are reliable in moral decision-making is problematic. As I formulated it, the normative challenge consists in understanding whether we are justified to infer normative conclusions from an increased understanding of the processes underlying moral judgments and decisions[127]. A more detailed description of these processes, therefore, might be of help.

In the past decades, dual-process frameworks have been characterized in several ways: fast vs. slow, automatic vs. controlled, unconscious vs. conscious, habitual vs. goal-oriented, affective vs. rational. I will focus here on a dual-process framework for morality which I believe to be more promising than others for several reasons. First of all, this framework denies the problematic – though extremely common and influential – emotion/reason divide. Although this distinction has (historically) been a favorite way of philosophers to understand moral psychology, both critics and advocates of dual-process models have recognized that positing a clear distinction between emotions and reason (or affective and "cognitive" processes) is incorrect, since both type 1 and 2 processes always involve integrative information-processing as well as affective and motivational components (Saunders 2016).[128]

Denying the emotion/reason distinction, however, does not mean leaving *any* dual-process account of moral cognition behind. Experimental research shows that two types of processes can be distinguished in moral as well as in non-moral decision-making, although framed in different ways, and portrayed as deeply interacting and cooperating.

A promising strand of dual-process models (Crockett 2013; Cushman 2013), relatively under-considered in the philosophical literature, frames moral cognition by stressing the distinction between:

1) Attributing value *directly to actions* by associating positive or negative value to them on the basis of a history of feedback (e.g. rewards or losses);
2) Attributing value to expected *outcomes* on the basis of a causal model (a "cognitive map") representing options, values, and transition functions.

These frameworks have two immediate advantages. First, they account for the presence of affective and cognitive information-processing in both types of processes; second, their reliance on learning models account for the diachronic dimension of moral cognition

[127] Note that the same strategy is adopted by those who defend the higher reliability of type 1 processes: since they can learn and be sensitive to reasons – they argue – type 1 processes can be reliable.
[128] For instance, processes leading to C judgments do not just elaborate the factual information "5 is more than 1", but also affective elements leading to endorse, or choose, that "saving 5 lives is better than saving 1". Moreover, both D and C judgments involve factual information processing: D judgments and emotional reactions are always driven by a clear representation of structural features of the situation, such as personal interaction, the exercise of bodily force (Greene et al. 2009), or direct vs. indirect harm (Royzman & Baron 2002; Cushman et al. 2006).

significantly more than first-wave dual-process models did.

These models are also consistent with several studies in moral psychology reporting a preference for indirect over direct harm (Rozyman & Baron 2002), strong aversion to typically harmful actions even when fake or victimless (Cushman et al. 2012a; Haidt et al. 1993), and the systematic presence of moral norms across history and societies prescribing the wrongness of specific action-types independently of outcomes (e.g. rituals, food and sexual taboos) (see Graybiel 2008). In these cases, characteristically deontological responses are elicited by the value directly associated with actions, regardless of other relevant information, such as expected outcomes or empathic concern for the subjects involved.

In addition to this evidence, action-outcome frameworks are supported by recent research in computer science and computational neuroscience, reflecting the difference between two basic kinds of reinforcement learning: *model-free* and *model-based* algorithms (Dolan & Dayan 2013).

Model-free learning and decision-making

Model-free (MF) algorithms work by associating positive or negative value to specific and immediately available actions after a history of rewards, independently of a causal representation of the environment. Imagine an agent A who, when turning right in a state *r* (*round*), gets a reward. If this association occurs a significant number of times, A will associate a positive value to the option "turn right" when in *r* states. Now imagine that A reaches state *r* after turning left in a state *s* (*squared*). Since A associates positive value to state *r*, A will also associate positive value to the option "turn left" when in *s*; and so on, creating adaptive chains of actions.

This mechanism brings A to associate value to the available actions in each particular state on the track leading to a reward, treating each of them as if it was itself a reward. The main advantage of this algorithm is that it is computationally cheap: at each step, it decides on the basis of the value associated with the immediately available action, avoiding costly simulations of future or hypothetical states and comparisons between them. However, and precisely for this reason, MF algorithms are not farsighted. They cannot be goal-oriented – nor prospective in general – because they lack a causal representation of the relation between possible actions and outcomes. This precludes them from any chance to make plans at all: MF algorithms are fundamentally retrospective.

Moreover, although very efficient, MF algorithms are inflexible. They cannot use

information to adjust values associated with states, actions, and outcomes (and, consequently, preferences and behavior) because they lack a global representation of them. Value representations can be updated, but this requires time, trial-and-error learning, or interference of strong opposing values (Dickinson et al. 1995).

Model-based learning and decision-making

By contrast, model-based (MB) algorithms choose by considering available courses of action on the basis of a causal representation – a model, or a cognitive map – of the environment. The model includes causal relations between events (actions, outcomes, rewards, and transition functions) to which A attributes different values; the expected values of the available options are compared, and choices are taken by exploring the decision tree and via CBAs (Dolan & Dayan 2013).

The main downside of this algorithm are its computational costs. Nonetheless, MB strategies can be very flexible, because the model can be updated at any moment by integrating new information and changes in the environment. Imagine that agent A has identified the optimal strategy to reach a reward. Knowing that an obstacle is obstructing the optimal policy (e.g. the fastest route) can make A choose the preferred alternative option in the most efficient way (e.g. without having to face the obstacle on the fastest route before finding an alternative). MB algorithms can be very farsighted, because they can identify clear and complex policies made of long chains of actions, simulating and evaluating consequences of consequences, and modulating value representation accordingly.

In human (moral) cognition, these two types of algorithms interact deeply (Cushman & Morris 2015; Kool et al. 2018). MF mechanisms do not only regulate motor habits or personal harm-aversion, but also the application of rules, principles, and concepts (Dayan 2012); they also facilitate MB decision-making by proposing limited sets of possibilities, thus avoiding the consideration of potentially infinite options in deliberative planning (Phillips & Cushman 2017). But to what extent can the differences between these algorithms – and/or their interaction – be normatively significant?

2. Normative relevance

Greene (2017) argued that the MF-MB distinction provides further support for

consequentialism.[129] Like fast-and-frugal heuristics, MF decision-making is generally reliable in front of ordinary contexts and problems, but "it would be a cognitive miracle if we had reliably good moral instincts about unfamiliar moral problems" (Greene 2014, 715). If this makes sense, then more new, complex, and controversial moral problems require MB reasoning. Since empirical research shows strong correlations and similarities between MB thinking and consequentialism (Patil et al. 2020; van Honk et al. 2022), Greene concludes that the latter is the best normative theory to address those kinds of problems.

Note that according to Greene – as for many other advocates of consequentialism – this does not mean that agents should engage in CBA *all the time* (Hare 1981; Brink 1989). MF decision-making can work well in many circumstances, but MB reasoning is more reliable when we have to decide about complex cases, as well as about moral principles, rules, procedures, decision strategies, and whether or not to trust our intuitions. Advocates of deontological and virtue theories, Greene argues, deny this, favoring forms of MF thinking such as reliance on norms or the moral perception of virtuous agents.

These conclusions are partly convincing, but also partly problematic. On the one hand, Greene addresses the normative challenge in a promising way. Consider the following characterization that Railton (2017) recently gave of moral inquiry. Unlike other domains (but similarly to science) the moral discourse aspires to overcome subjective, tribal, elitist, or esoteric points of view and interests by following procedures, and looking for understanding and justification that are impartial, general, consistent, authority-independent, shareable, thinking- and action-guiding, and non-instrumentally concerned with interests and reasons of those actually or potentially affected (Railton 2017, 173).If this characterization is plausible, then the only decision strategy able to accomplish these tasks cannot but be MB reasoning. Consistency, for instance, would be impossible without a model representing the value associated with principles, actions, and outcomes. MB reasoning is also the only strategy allowing us to consider the interests and reasons of others beyond our natural and cultural inclinations, and to evaluate them critically in light of relevant information and alternative possibilities. Moreover, consistent and intersubjectively acceptable moral justifications (Songhorian et al. 2022) cannot but be MB. Referring to a model – models are non-perspectival by definition – is the only way to make one's reasons intelligible to others. Finally, MB

[129] Greene (2014) illustrates this idea through the analogy with a camera's automatic vs. manual settings. As he noticed later, however, this analogy can be misleading because the automatic settings of standard cameras do not change after they leave the factory, whereas "people's "automatic settings" are constantly evolving through learning [...] The key point, however, is that at the time of decision one is stuck with the automatic settings that one has, regardless of how circumstances might have changed" (Greene 2017, 5).

reasoning is necessary to link immediately available actions with distant goals, and to consider alternative courses of action (Railton 2017).

On the other hand, however, the idea that the higher reliability of MB reasoning supports consequentialism is problematic. The empirical literature is partly inconsistent on this matter; there are, nonetheless, at least four reasons to doubt such a bold normative conclusion.

1. Studies on confidence and decision-time in moral decision-making suggest that non-C judgments might be the result of MB reasoning *also at the time of decision* (Koop 2013; Gürcay & Baron 2017; Bialek & De Neys 2017)[130];
2. MB reasoning should not be identified uniquely with CBA in act-utilitarian terms, but rather as a broader reflective operation that considers i) information, potential courses of actions and outcomes, ii) intuitions, feelings, rules and principles, and iii) reasons, testing their reciprocal consistency and discarding recalcitrant options (Brink 1989; Campbell & Kumar 2012; Bazerman & Greene 2010).
3. D/non-C judgments can be justifiable even when they are the proximate output of MF processes. First of all, they can be the (distal) output of previous MB reasoning or rationalization. In some cases, justificatory reasons can even track some processes that led to the new "educated" intuition, even if these processes did not intervene at the time of decision (Sauer 2017; Kumar 2017).
4. Finally, also C judgments can be the result of MF processes (Bago & De Neys 2019). For instance, Trémolière and Bonnefon (2014) have shown that the higher the number of lives involved in sacrificial dilemmas, the more intuitive C judgments are. This suggests that C responses can be model-free too, requiring MB reasoning when they are more counterintuitive (Kahane 2012).

To sum up, empirical research and the MF-MB framework support important normative conclusions, though mostly in 'procedural' terms, i.e. suggesting how we should think in front of complex or new decisions, and how to justify them. This, however, has no clear direct implications for normative ethical theory in a more substantive way.

Some readers might still be unconvinced about the procedural normative conclusion that MB moral reasoning is more reliable than MF mechanisms to address new and complex moral

[130] It is worth noting that Greene recently changed his mind because of this literature. Greene still stands by the general dual-process story, but no longer thinks that MF is literally faster than MB (at least on the time scale of reading and responding to moral dilemmas). But he still thinks that alignment of MF with D, and MB with C (in classic dilemma contexts) holds, as he considers the evidence for this – especially from lesion studies – very strong. A very brief sketch of this revised view can be found in Greene (forthcoming).

problems. I will briefly consider two possible reasons in favor of this skepticism:

i) In a recent paper, Cecchini argued that default-interventionist models of moral cognition – according to which type 2 (MB) processes intervene to control, endorse, or reject type 1 (MF) outputs – are inaccurate because (MB) moral reflection *fundamentally depends* on (MF) intuitions (Cecchini 2021, 301). In fact, recent research suggests that:

i.i) MF mechanisms often *facilitate* MB reasoning, providing by default limited sets of options within potentially infinite ones (Phillips & Cushman 2017);

i.ii) MF mechanisms *detect conflicts* between intuitions, reasons, and non-moral information, signaling the need for further reflection (De Neys 2014).

Although these claims are descriptively true, by no means they constitute an objection to the normative conclusion defended here. Operations such as cognitive filtering and conflict detection are not intrinsically reliable: they might be based on, and lead to, either reliable learning histories and actions, or biased and unjustifiable ones.[131]

Consider these two cases. First (i.i), agent A might not even consider being fair or kind to a member of a discriminated group, or engaging in sustainable behaviors, because these options might not be included in the default set provided by MF processes as a result of her learning history. Her habits are different and pretty inflexible; she can contemplate different possibilities, but she does not consider *those* actions since the value associated to them is significantly lower than alternatives available at the time of decision. Second (i.ii), intuitive conflict detection and resolution might result in discarding reasonable options (e.g. the less harmful, or the more supported by evidence) because too costly to hold; the conscious reasoning process called upon by intuitive conflict detection might be merely confirmatory of pre-reflective intuitions (Kunda 1990; Haidt 2001).

There is hence no reason to hold MF mechanisms trustworthy in the moral domain just because of their causal role: decisions are often driven by intuitive (MF) processes, but in no way this justifies them. On the contrary, the aforementioned limits of MF algorithms cast doubt on their outputs if no specific convergent support is provided by MB reasoning. In both the aforementioned cases, only MB strategies can critically evaluate whether to endorse the input provided by MF default options or to consider alternative ones. Moreover, only MB reasoning can test whether intuitions are reciprocally consistent and supported by reasons, independently

[131] In order to respect Railton's criteria for non-perspectival moral inquiry mentioned above – i.e. for being intersubjectively communicable, understandable and justifiable –, the normative standards needed to assess the reliability of cognitive processes and behavioral outputs cannot but be model-based.

of pre-reflective confidence about their rightness. Deciding uniquely based on the strength of 'feelings' or 'seemings' is not a defensible strategy (Brink 1989, chapter 5; Harris 2012, 294).

ii) Finally, MF-type 1 mechanisms have been indicated as responsible for the meta-cognitive task of deciding whether MF or MB strategies should be implemented to address specific problems (Cecchini 2021; Thompson et al. 2011)[132]. However, recent studies suggest that when facing a problem, people often engage in CBA weighing the expected outcomes of each strategy (including, in the calculation, the computational costs of MB reasoning), rather than relying on heuristics. Specifically, data show that engagement in MB reasoning – both as metacognitive arbitrator and as the ultimate decision strategy – is proportional to the stakes and levels of uncertainty involved (Kool et al. 2017, 2018). These results are consistent with previous research suggesting that at each time point agents estimate the expected costs and rewards from engaging in a full MB estimation of action-outcome values (Keramati et al. 2011). Although MF processes do play a role in this arbitration, there is no reason for holding them as reliable detectors of the right decision mode for specific and complex problems (Bazerman & Greene 2010).

In this chapter I argued that dual-process models of moral cognition are plausible, though they should not be framed in terms of the problematic emotion/reason dichotomy. I also suggested that the distinction between model-free and model-based learning and decision-making algorithms can lead us to draw important normative conclusions. Specifically, in light of a) how they function, and b) the problems we have to face, this framework supports the higher reliability of model-based moral decision-making in front of new, uncertain, and/or complex scenarios. Reliability can be conceived of in terms of justifiability: people would more likely provide – and freely accept – good moral justifications based on non-perspectival model-based reasons, rather than on the subjective 'feeling' or 'smell' of what is right (although this latter strategy that can give rise to effective *post-hoc* rationalizations; see Songhorian et al. 2022). These conclusions, however, are procedural rather than substantive. Indeed, model-based moral reasoning should not be seen as merely evaluating outcomes (Cushman 2013), nor as a kind of purely consequentialist form of thinking (Greene 2017), since it can be open to the consideration of several non-consequentialist reasons, norms, intuitions and evaluations (Białek & De Neys 2017). The coherentist mechanism needed to balance all these considerations is a form of model-based reasoning, though it looks closer to a reflective equilibrium than to a pure cost-benefit analysis.

[132] Evans (2019) hypothesizes a 'type 3' process for this task, presenting aspects of similarity with both type 1 and type 2 processes.

11. Moral reliability and expertise

In this final chapter, I move some criticisms to the idea of moral expertise. First, (1) I introduce the idea that moral expertise would require the possibility of acquiring objective and specific moral knowledge. In (2-3) I criticize this approach, together with the thesis according to which the acquisition of moral competence would take place in a similar way to the learning and development of skills in other fields. According to this thesis, moral experts have developed an intuitive competence based on their experience with moral values or problems, and this makes them authoritative and reliable judges as they are more able than others to provide appropriate answers to these kinds of challenges. I suggest that this thesis is problematic, also in light of the recent advances in empirical research on the cognitive processes involved in moral learning and decision-making discussed in the previous sections.

I propose an alternative approach for assessing different degrees of competence and reliability of moral judgments and judges that is less controversial and more easily operationalizable and implementable than the view that identifies moral experts based on more substantive criteria. According to this perspective, the reliability and authority of a moral judge in specific contexts is a function of the respect of a set of reliable procedural constraints, combined with the exercise of some of the virtuous character traits discussed in chapter 9.

1. Moral knowledge and expertise

According to moral realists, the recognition of moral expertise should depend on the possibility of acquiring objective moral knowledge, whether this knowledge is possessed to a greater extent by experts over non-experts, or whether it consists in the objective moral standard by which people can evaluate the reliability of alleged moral experts.[133] This thesis seems reasonable: as it happens in several other domains, even in order to identify competent and authoritative subjects in the case of moral problems it seems necessary to define or recognize the existence of objective normative standards (McGrath 2008). According to this view, moral experts would be those who, more than others, know or contribute to discovering what this standard consists of; who recognize its instances or violations in specific cases, and know how to inform, review and correctly translate into practice the rules and principles that define it.

Imagine, now, that Vera, a historian and anthropologist, knows in detail the moral norms of many human societies, past and present. Her specialist knowledge makes her an expert on these issues and phenomena. However, this is not the kind of knowledge that realists consider necessary for recognizing expertise and normative authority in morality. Vera's knowledge is just descriptive of facts or truths about the morals of specific populations, such as: "the moral principles, norms and judgments *x*, *y*, *z* are articulated in this way, and they are true for someone" (e.g. a specific social group). Vera's knowledge of such relative or contextual truths is something different from knowing the objective or stance-independent validity of their normative content. Vera is just an expert on what certain people or societies believe is right or wrong, not on what is right or wrong *per se*.[134]

According to moral realists, to know *x*, *y*, *z* means to believe that *x, y, z* are objectively true (Shafer-Landau 2003). Many people do not know the objective validity of these truths, nor what justifies them; experts, on the other hand, understand them clearly, distinctly and coherently, recognize or perceive them in situations, and/or act in accordance with them on a regular basis.[135] In what follows, I argue that this approach is problematic.

---

[133] Brink (1989, IV.7), Shafer-Landau (2003). According to non-naturalist moral realism, this kind of knowledge is also specifically moral: moral facts are neither natural nor supernatural, but *sui generis* (Broad 1930; Moore 1903/1993; Prichard 1949; Ross 1930/2002).

[134] Nobody denies that it is possible to know facts or truths concerning morality as a psychological, social, or bio-cultural phenomenon. Anti-realists deny that there are intrinsically normative, or 'first-order' moral truths. The proposition that these kinds of facts do not exist is true for the anti-realist, but this is a 'second-order' truth about morality (meta-ethical, not normative). See Mackie (1977), Greene (2002, 7).

[135] Some scholars have conceived of moral expertise in a strong intellectualist, theoretical sense (Singer 1972; Singer & Wells 1984; Crosthwaite 1995); others, in line with the skill-model introduced in chapter 9, in a much more performative sense (Annas 2011; Dreyfus & Dreyfus 1991; Stichter 2018)

2. Objective standards and disagreement

How can we identify moral experts? In many contexts, agents are recognized as experts on an inductive basis, by virtue of their track records by reference to relatively uncontroversial normative standards, or in light of other social (e.g. reputational) indicators of competence, authority, and trustworthiness. In the moral sphere, this operation is more problematic. Following an analogous approach, in order to identify who knows, advises or makes the most appropriate choices in most cases, we would need to dispose of shared moral standards; often, however, there is no agreement on what the correct or best normative standards are, nor which decisions are more morally appropriate in specific cases – even in light of a relative agreement at the level of more abstract and general principles (see also section 3.1 on this point).

Doctors, researchers, athletes, cooks, musicians, journalists, sales assistants can all improve their technical skills and performances and acquire expertise in their field in response to several forms of environmental and social feedback. For instance, behaviors and compliance with norms or principles that lead to errors and/or incur in high costs, damages, or blame are subsequently avoided or revised; those that receive positive feedback are reinforced. Similar dynamics also occur in moral learning: even prosocial behaviors or moral reasoning that involve high costs are often reinforced by positive feedback effects for those who manifest them (Railton 2017).

However, it is possible to distinguish between the main dynamics and the effectiveness of learning processes, on the one hand, and their compliance (or not) with normative standards on the other. As introduced in chapter 1 (and also discussed in relation to moral development in chapter 9), learning is a neutral process, not an inherently positive one; it can be evaluated differently, from a moral point of view, depending on cases and reference standards. It is one thing to observe or describe the socio-psychological dynamics of development, learning, and moral change; it is another to make evaluative judgments about their goodness or desirability, or whether and to what extent they constitute improvements or progress. As noted by Antti Kauppinen,

> If you judge that abortion is wrong even if it is not and act on your belief, there is no negative feedback that results simply from your having made a moral mistake. (The only reliable negative feedback you will get for acting on a moral judgment is from people who disagree with you, but that is not an indication that you are wrong.) So we cannot train our intuitive system to respond to moral truths in the same way we can train it to respond to truths about good chess moves or ill infants. The expertise defense of moral intuitions is unsuccessful (Kauppinen 2014, 295).

Certain issues appear to be characterized by a greater degree of normative uncertainty ("what would be appropriate to do") and genuine disagreement than others. Quoting Sidgwick (1907/1981, 342), Sarah McGrath proposes to consider a belief *controversial* "if and only if it is denied by another person of whom it is true that: you have no reason to think that she or he is more in error than you" (McGrath 2008, 91). Moral beliefs and judgments are often controversial in this sense. In these cases, there is good reason to doubt that our opinions constitute objective, justified knowledge. If it is not possible to claim to possess knowledge about certain issues, we lack the objective standard that the realist thesis considers necessary for the recognition and evaluation of moral experts.

A natural objection to the epistemological argument just proposed is that disagreement is not enough to doubt our knowledge of something: the fact that many deny it does not make the theory of evolution a controversial opinion. But evolutionism is not controversial precisely because we have reasons to believe that those who deny it are wrong, or that they do not base their conclusions on epistemically reliable methods (e.g. ignoring the evidence or the mechanisms of production of scientific knowledge) (ibid., 89, 96).

However, various moral problems seem controversial even among people who share comparable levels of relevant empirical knowledge and familiarity with analytical procedures and tools such as moral theories. In these circumstances, it is difficult to identify who, better than others, knows the correct answer or stance to take in substantive terms.

3. Procedural moral improvement and moral justification [136]

Should we conclude that all judgments and judges are equally reliable in the face of complex moral problems? No: evaluating moral judges as more or less reliable than others is possible even without postulating the existence of objective moral standards as realists do. As already stressed in previous sections of this work, I suggest we shift our focus to the processes and methods that lead to the formulation of judgments, decisions and moral justifications, rather than to their substantive normative content.

The tools we can use to assess the processes that lead to certain judgments, actions, and instances of social change are disparate, and will likely grow and improve in accuracy in the

[136] A part of this section has been re-adapted from Songhorian et al. (2022, sections 3 and 4, 179-185). I am grateful to Francesca Guma, Massimo Reichlin, and especially to Sarah Songhorian for their contribution and important discussions on these issues.

future. In Parts I and II, I suggested that assessing the degree of agency driving judgments, beliefs, and behaviors seems a justified proxy and feasible project to partly assess their moral worth (and 'progressiveness' in historical evaluations). Above in Part III, I have pointed out how in recent decades, developments in experimental research in cognitive and computational sciences have contributed to providing a sort of

> "behind the scenes" look at human morality. Just as a well-researched biography can, depending on what it reveals, boost or deflate one's esteem for its subject, the scientific investigation of human morality can help us to understand human moral nature, and in so doing change our opinion of it (Greene 2003, 847).

Here, I will limit myself to a brief discussion of some possible implications of these ideas for the debate on individual moral improvement and moral expertise.

Over the course of the chapters, we came to appreciate the fact that increases in agency and decisional autonomy are closely related to emancipative processes from egocentric, ultra-partialist, parochial, and biased perspectives, preferences and values, as well as from other strong normative constraints, such as biological pressures, habits, fear, deference to authority and tradition, ignorance, and so forth. But if the most reliable decision-making strategy for tackling controversial moral issues is to engage in model-based reasoning over intuitive, model-free responses, to provide non-perspectival, coherent justification and be open to understand those of others, then moral experts may be those who, better than others, develop and exercise the domain-general skills and competencies which are needed for these operations. In this sense, the possibility of evaluating differences between moral judges in terms of reliability and authority should be understood in procedural rather than substantive terms.

On this account, we should not look only at people's actual behavior nor at the content of their moral judgments – although both are certainly relevant –, but rather at the abilities and procedures needed to better understand moral problems and justify our conclusions or responses (Schaefer & Savulescu 2019; Rawls 1951). What is relevant in this perspective is not what individuals do, judge, or believe, but rather processes with which they reach their conclusions, and the reasons and justifications they can provide in support of their actions, judgments, and beliefs. Consistent with my agency-based theory of moral progress presented in Part I, a considerable criterion to assess the reliability of judgments and choices is to evaluate how a moral output is reached and justified, rather than to focus only on what that output is.

In a similar fashion, Schaefer and Savulescu (2019) recently provided a Rawlsian-inspired set of procedural features to assess the reliability of moral judges. These features are:

• *Logical competence*, i.e. moral judgments should be mutually coherent: respecting this constraint requires the capacity to make correct logical inferences, spot inconsistencies in one's and others' judgments, identify the implications of one's beliefs and the matter of contention between interlocutors.[137]

• *Conceptual understanding*, i.e., deepened understanding of the content and scope of application of moral concepts and ideas, and the ability to communicate it effectively (77; see also Moody-Adams 1999).

• *Empirical competence*, i.e., knowledge of non-moral, empirical facts[138]

• *Openness to the revision* of one's opinions (see chapter 9 above);

• *Empathic understanding*, i.e. understanding of others' situation (79);

• *Bias avoidance*, i.e. "taking factors into account in a moral judgment that are not relevant to that moral judgment"[139] (see also Greene 2014).

I will to discuss and justify each of these requirements here (for discussion, see Schaefer and Savulescu 2019), and this list should be understood as partial and open to revision. But I think this constitutes a promising example of how one can and should proceed.

Two observations, however, deserve to be made. First, some of these features can also be understood in terms of intellectual virtues (e.g., openness to revision as epistemic humility) and, as outlined in chapter 9, virtue theory can offer several tools to improve their development and facilitate the respect of some of these requirements. The second is a link with the discussion about the functional specificity of moral cognition and the evolution of open-ended moral

[137] "One might hold, for instance, the following three views: all corrupt politicians should be punished no matter how mild the corruption; one's favourite politician is mildly corrupt; and one's favourite politician should not be punished for so mild a corruption, given all the good work she is doing. These are jointly inconsistent, as the first two views imply by modus ponens that one's favourite politician should be punished even for mild corruption. Something has to give – logically, one of the views must be given up" (ibid., 76; on this point, see also Brink 1989; Campbell & Kumar 2012; 2013).

[138] Consider this argument proposed as an example by Schaefer and Savulescu:

"P1: Senator Barney accepts bribes

P2: Anyone accepting bribes should be punished

C: Senator Barney should be punished

P2 and the conclusion are moral claims, and so without further elaboration are untouched by empirical concerns. However, P1 is an empirical, non-moral claim. The moral conclusion only follows if it is correct. Anyone endorsing the conclusion that Senator Barney should be punished on the basis of the above reasoning needs to have good grounds for the claim that Senator Barney accepts bribes. Some sort of evidence such as a witness of the bribery will be needed. And those evaluating such evidence will need to assess a number of factors. Is the witness reliable? How do we know what was witnessed was really a bribe? What did the briber procure? Those who are generally more competent at evaluating empirical claims will more reliably ascertain the truth of P1, and in turn make more reliable evaluations of the moral question of whether Senator Barney should be punished" (77).

[139] E.g. "how you frame a question should not matter to one's opinion of it; one should not hold oneself to different moral standards as that of others; one should not privilege one's relations over others in the public sphere; and so on" (81).

reasoning carried out in Part II. The combination of improvements in intellectual and epistemic virtues and in the respect of these procedural constraints seems very far from being an improvement in one, domain-specific psychological capacity dedicated to morality. It looks, rather, much more like a complex combination of improvements in several other abilities.

There are at least two reasons for a procedural account like this to be preferable over more substantive ones. First, it enables one to hold a pluralistic stance "thus avoiding many question-begging moral assumptions" (Schaefer & Savulescu 2019, 75). Several moral disputes are, in fact, so controversial that it is problematic to believe that one solution is the true or correct one, that everyone has reasons to accept. Second, a procedural account – especially one that is concerned primarily with how people justify their behaviors, decisions, judgments, and beliefs – is more suited to account for instances in which an individual might have come to a moral conclusion because of external or internal drives that would not count as an appropriate moral justification of that conclusion (rather, it would cast doubt on it). Let us now delve deeper into these two issues.

As far as the latter is concerned, to say that individual moral improvement only consists in performing – or complying with – practices judged by a third party as 'morally better' overlooks the possibility that behavior can be influenced or causally determined by manipulation or indoctrination. How can we distinguish between someone who is getting rid of her biased behavior towards a social group because she has understood that it was grounded on faulty bases (so that she now believes it is morally wrong to have that behavior to begin with), from someone else who does the same just because it is fashionable to be seen as open-minded?

In this case, the behavior change will be relevant to account for a person's improvement, but it will not be sufficient. Indeed, it is difficult to say whether a change is determined by an effective, stable, and authentic moral improvement by only observing behavior: people could act in a certain way because they are influenced by internal or external stimuli, by their desire to be socially approved rather than by that of deserving approbation (Smith 1759/2004, III.2.32), by morally irrelevant factors rather than by the morally pivotal ones. On the assumption that an authentic moral action involves a strong sense of agency of the subjects, focusing on how judgments are made and on how moral behavior is grounded can reveal a way to increase the agents' real moral capacity and the conscientiousness of their moral responses (Schaefer 2015).

Coming to the first issue, the adoption of a pluralistic stance points clearly away from accounts measuring moral change and moral improvement only in terms of their behavioral

outputs. There are many contexts of choice where the issues involved are so disputable, and/or where no action is clearly recommended, that believing one particular behavior represents the right way to go means assuming a specific normative outlook, one that might not be universally shared. Is there a clear set of actions that we can universally conceive as the right one when dealing with issues such as, say, the scarcity of healthcare resources or global poverty? Since the answer appears negative, it seems reasonable to focus on how people justify their often-divergent beliefs and behaviors; endowing people with a sensitivity for the reasons at stake and a capacity to respond to them helps reduce moral plurality by excluding those moral stances that do not pass the test of justification. Thus, improving the abilities and faculties that are involved in an appropriate moral justification should be the starting point to promote moral improvement and moral change.

By aiming to avoid the imposition of a substantive normative standpoint as the only right or best one, a procedural account lowers the risk of indoctrination, manipulation, and paternalism in the promotion and assessment of moral improvement and aims to track the path to enhancing moral agency. Focusing on individuals' abilities to provide reasons according to logical, empirical, and conceptual competence, openness to the revision of one's opinions, sympathetic imagination, and bias reduction – i.e., the abilities Schaefer and Savulescu focus on – is a good starting point to ascertain whether one is truly improving her moral stance. Thus, while a behavior or a judgment for which the subject can provide (convincing) reasons is certainly better than one for which no justification seems to be available to her, this clearly is not the end of the story nor a solution for every moral dispute. Much is still lacking for a complete account of individual moral improvement to be in place.

To pave the way for it, though, a procedural account like the one I have gestured towards here is required. In the remainder of this section, I will consider a challenging objection to such an account: how can we be sure that improving someone's ability to provide reasons for her actions leads to a moral progress and not a regress? How can we be sure that a procedural account of moral justification has the resources to distinguish proper justification (or moral reasoning) from vicious post-hoc rationalization or confabulation (Haidt 2001; Greene 2007)?

As mentioned, by avoiding any substantive commitment, this proposal risks considering an amelioration in the formal ability to rationalize any moral (or immoral) conclusion as a proper instance of moral improvement. I offer some replies to this concern by suggesting that not every reason-giving account counts as a proper moral justification, and by adding some considerations about the empirical and theoretical assumptions which may ground this worry.

This objection may stem from views sympathetic to Haidt's influential model of moral

judgment (Haidt 2001). Drawing on empirical research, Haidt concludes that moral judgment is not the product of conscious reasoning, but the expression of automatic, unconscious, and affectively-laden "intuitions" shaped by evolutionary, cultural, and social pressures. Within this model, conscious reasoning intervenes only ex post by concocting reasons to support and socially justify fast and automatic reactions: "one feels a quick flash of revulsion […] and knows intuitively that something is wrong. Then, when faced with a social demand for a verbal justification, one becomes a lawyer trying to build a case rather than a judge searching for the truth" (Haidt 2001, 182). According to Haidt, the function of moral reasoning is to socially justify intuitive responses, but it has no power in shaping their content ex ante. In this framework, increased proficiency in the ability to provide socially acceptable justifications would just better perform the function of convincing others about the acceptability of conclusions that are essentially insensitive to rational scrutiny and revision.

I believe that satisfying certain procedural requirements can make certain reasons or justifications more intersubjectively acceptable than others, without committing to any substantive normative or metaethical view. In particular, some justifications can be more consistent, more sensitive to empirical evidence and to others' perspectives, reasons and interests, and more open to revision than others. Acceptable moral justifications do not merely confirm one's opinions, intuitions, or feelings by effectively convincing other people about their soundness; they also express the effort of considering a broader spectrum of information, such as non-moral facts, or the interests and preferences of the individuals involved (including the agent's ones).

If Schaefer and Savulescu's criteria are reasonable and sensible, one can discriminate between different levels of reliability or appropriateness of moral justifications, distinguishing between confabulations (and the correlative phenomenon of moral dumbfounding), motivated or confirmatory rationalizations, and appropriate moral justifications.

According to Haidt's work, a confabulation is the attempt to fabricate justifications for moral conclusions with clear fallacious results (e.g., blatant logical contradictions), pushed by the desire to hold and confirm one's feelings, intuitive judgments, and beliefs, even when put in front of inconsistencies and contrasting rational arguments (Festinger 1957; Kunda 1990). In Haidt's famous experiments, some subjects try to rustle up support for their intuitive conclusions by offering fallacious and unsatisfactory justifications which, for example, patently clash with relevant information or just restate intuitive conclusions without justifying them at all (Haidt 2001; 2012). Therefore, we can conceive of confabulation as a vicious kind of reason-giving, which lacks several features of an acceptable justification (such as empirical

and logical consistency and openness to revision).

Rationalization can be understood, more broadly, as the justification of behavioral outputs by offering reasons in their support "that would have made it rational" (Cushman 2020, 183), even if such reasons do not match the actual processes that led to that output. Many rationalizations can be more consistent and sensitive to logical reasoning and evidence than moral confabulation. However, providing reasons in favor of a moral judgment does not guarantee providing acceptable moral reasons because what is rational, for instance, from a self-interested point of view may not be so from a moral point of view. For example, a rationalization may be grounded on an astute selection of data, aimed to make the preferred conclusion plausible, while a proper moral justification does consider more morally relevant factors, such as the interests of other individuals involved. Also, while rationalization does not require critically examining one's own moral preferences, a good moral justification does. Moreover, even though rationalization requires paying attention to possible influences of biases on argumentation, it does not require taking seriously, for instance, the main moral reasons for and against available stances or lines of action. Supporting moral conclusions with acceptable moral justifications does not simply require a generic capacity to provide any kind of reasons in their favor, but to provide a much more specific kind of reason-giving account.

An acceptable moral justification, thus, requires adequately knowing the context of the situation under evaluation, along with one's and others' perspectives. Reasons for and against different conclusions should be balanced in light of available information, showing the attitude to evaluate potential alternatives with an open mind, and being disposed to reconsider one's opinions. The potential influences of biases or prejudices that might affect the evaluation should also be considered. To achieve this goal, it is important to avoid considering one's preferences as the right evaluative standard for the situation at hand, acknowledging and balancing the different interests at stake. Finally, acceptable moral justifications should satisfy standards of logical consistency. Improvement in these capacities does not merely enhance the formal ability to justify any possible moral judgment or behavior – as the objection we are addressing states – because if these requirements are satisfied the spectrum of reasonably acceptable moral conclusions shrinks significantly.

Note that a strength of this view is that it stands even if Haidt's model of moral judgment is plausible. Even if in isolated, specific circumstances of choice explicit moral reasoning intervenes only after quicker psychological responses, improved justificatory abilities would not just better support a-rational outputs, but can be sensitive to independent relevant information. Nonetheless, there are several reasons to reject Haidt's thesis according to which

moral reasoning has no causal influence on moral feelings, intuitions, and judgments. Several critics have stressed the limits of Haidt's model, denouncing its rigid lack of interaction between controlled and automatic processes, as well as its blindness about the diachronic dimension of moral judgment (Campbell & Kumar 2012; Railton 2014). Even if it does not come into play immediately before the expression of a moral conclusion at the time of decision, explicit moral reasoning can feedback on, inform, and improve people's future moral responses (Sauer 2017). If this is true, an appropriate moral justification can also reliably point out some of the reasons that informed one's intuitive judgment or behavior (Cushman 2020). All these are not necessary requirements of motivated (or confirmatory) rationalizations. Therefore, I suggest that acceptable moral justifications can be distinguished from other reason-giving accounts. This allows me to reject the objection accusing my procedural view of considering mere improvements in the capacity to rationalize as proper moral improvements.

In conclusion, back to reliability and expertise, by evaluating the level of satisfaction of each of the parameters discussed above, the procedural method suggested here appears relatively easy to operationalize for the identification of the most reliable judges in specific contexts, as well as for the implementation and measurement of improvement projects. While morality, agency, and open-ended normativity are nothing supernatural, and the cognitive mechanisms and psychological capacities involved in them are likely the same recruited for other kinds of cognition, behaviors, and domains, the analysis of ethical problems, moral reasoning and moral justification can be favored by specific virtues and by the respect for certain procedural constraints. Respecting those criteria – including the very possibility of revising them – and understanding how to favor the development of traits of open-mindedness and mental flexibility looks like a viable path to progress, and to its understanding.

Conclusions

Throughout the course of this work, I touched on and explored several problems and attempted to offer my perspective on them. I have no ambition of having solved any of them, but I hope I have been able to put forward a few insights on which to build, possibly, future research and projects.

In Part I, I suggested the importance of reflecting on the normative-evaluative core of the idea of moral progress, a project which has been underestimated and even openly rejected in recent discussion about this topic. I suggested a working 'dual' agency-based account of moral progress, according to which increases in agency and decisional autonomy seem to be a fairly convincing and flexible proxy to assess and justify the 'progressiveness' of instances of moral and social change.

In Part II, I argued that a commonly held thesis in the contemporary scientific and empirically informed philosophical debate – the idea that human psychology remained unaltered since the Pleistocene – is incorrect. I supported this conclusion by relying on critiques of influential evolutionary explanations of moral cognition, and by discussing recent cultural-evolutionary hypotheses and cross-cultural empirical evidence on psychological flexibility and value change. In the last chapter of Part II, I suggested that important aspects of human morality and moral cognition – above all, open-ended moral reasoning and normativity – should be conceived of as cultural byproducts of the selection of several (domain-general and domain-specific) psychological traits which can allow for their development and increased exercise.

The selection and development of psychological capacities which can enable increased levels of agency and decisional autonomy depend on favorable ecological and socio-cultural conditions, but the more agency increases, the greater is the causal power of agency on the creation of social institutions that respect and enhance it: agency-based moral change feeds itself.

As we unfortunately but undoubtedly experience, moral progress is not guaranteed. But my claim is that individual psychology and agency can play a relevant part in this picture. This suggests two things. First, efforts to improve human agency with 'ordinary' socio-cultural means (i.e. with no need of moral-bioenhancements) seems relevant and feasible for guiding agency-based moral progress, even at the broader, socio-institutional level. Second, this calls for a better understanding not only of the ecological, socio-structural, and developmental enabling conditions allowing people and institutions to develop and promote greater agency and autonomy, but also of the ethical constraints that should regulate their exercise.

In Part III, I offered a perspective on the idea and possibility of 'individual moral improvement', still conceiving of it as an integral component of my agency-based theory of moral progress. I argued that a virtue-based account is a good candidate, since virtues can play a fundamental role in facilitating and regulating the exercise of agency and decisional autonomy. In particular, emphasizing the role of intellectual virtues such as mental flexibility and epistemic humility provides a fruitful synthesis between stability and open-endedness. A virtue-based account of moral improvement seems therefore needed in an agency-based theory of moral progress such as mine, but it also does not seem to provide a sufficiently clear normative criterion to understand in which direction individual moral progress should be headed, nor how to address specific, complex and controversial moral problems. I supported this conclusion by referring to recent experimental research on moral psychology and computational models of learning and decision-making, which show the limits of affective learning and moral intuitions even if 'educated' or 'trained'. I suggested that our increased scientific understanding of the cognitive processes involved in moral decisions can have some relevant normative implications, though this does necessarily lead us to recognize the superiority of a specific normative ethical theory: these normative implications can be conceived of as procedural, rather than substantive. Finally, I considered some implications of these conclusions for the problem of assessing improvements and/or different levels of reliability in moral decision-making. I emphasized, again, several good reasons for preferring a procedural approach, and considered some implications of adopting such a framework for the controversial idea of moral expertise.

Much work still needs to be done on these issues, as well as on some of those that I only briefly tackled throughout the chapters and/or did not sufficiently or convincingly analyzed. While psychological change and epistemic improvement are possible, biases permeate our thinking, and sometimes we become too attached to our ideas. Reliable information, dialogue, and the understanding of other perspectives and reasons are often good ways to reduce some of these biases, and to challenge our unjustified intuitions and beliefs. Several opinions, arguments, and conclusions that I tried to convey in this work may likely be confused, unjustified, or wrong. Hopefully, further discussion, investigation, and experience will contribute to make my view on these complex and challenging issues clearer.

References

Aarøe, L., Petersen, M. B., Arceneaux, K. (2017). The behavioral immune system shapes political intuitions: Why and how individual differences in disgust sensitivity underlie opposition to immigration. American Political Science Review, 111(2), 277-294.

Acemoglu, D., Robinson, J. A. (2012). Why nations fail: The origins of power, prosperity and poverty. Crown Books.

Alexander, R. D. (1987/2017). The biology of moral systems. Routledge.

Alfano, M. (2013). Identifying and defending the hard core of virtue ethics. Journal of Philosophical Research 38, 233–260.

Algoe, S. B., Haidt, J. (2009). Witnessing excellence in action: The 'other-praising' emotions of elevation, gratitude, and admiration. Journal of Positive Psychology 4(2), 105–127.

Allport, G. (1954). The nature of prejudice. Addison-Wesley.

Anderson, E. (2010). The imperative of integration. Princeton University Press.

Andrews, K. (2020). The animal mind: An introduction to the philosophy of animal cognition. Routledge.

Annas, J. (2011). Intelligent virtue. Oxford University Press.

Annas, J., Narvaez, D., Snow, N. E. (eds.) (2016). Developing the virtues: Integrating perspectives. Oxford University Press.

Archer, M. S. (1995). Realist social theory: The morphogenetic approach. Cambridge University Press.

Aristotle (1999). *Nicomachean Ethics*. Terence H. Irwin (ed./trans.). Hackett.

Arnhart, L. (2005). Darwinian conservatism. Imprint Academic.

Arvan, M. (2021). Morality as an evolutionary exaptation. In J. De Smedt, H. De Cruz (eds.). Empirically Engaged Evolutionary Ethics. Springer, 89-109.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological Monographs: General and Applied, 70(9), 1-70.

Asma, S. T. (2012). Against fairness. University of Chicago Press.

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. Proceedings of the National Academy of Sciences, 117(5), 2332-2337.

Axelrod, R. (1984). The evolution of cooperation. Basic Books.

Ayala, F. (2010). The difference of being human: Morality. Proceedings of the National Academy of Sciences, 107(2), 9015–9022.

Bago, B., De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. Journal of Experimental Psychology: General, 148(10), 1782.

Baehr, J. (2011). The inquiring mind: On intellectual virtues and virtue epistemology. Oxford University Press.

Baehr, J. (2021). Deep in thought: A practical guide to teaching for intellectual virtues. Harvard Education Press.

Ballantyne, N. (2019). Epistemic trespassing. Mind, 128, 510, 367-395.

Banaji, M. R., Greenwald, A. G. (2013). Blind spot: Hidden biases of good people. Delacorte.

Barkow, J. H., Cosmides, L., Tooby, J. (eds.). (1995). The adapted mind: Evolutionary psychology and the generation of culture. Oxford University Press.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, A., Scelza, B. A., Stich, S., von Rueden, C., Zhao, W., Laurence, S. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. Proceedings of the National Academy of Sciences, 113(17), 4688-4693.

Baumard, N., André, J. B., Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. Behavioral and Brain Sciences, 36(1), 59-78.

Baumeister, R. F., Masicampo E. J., DeWall C. N. (2009). Prosocial benefits of feeling free. Personality and Social Psychology Bulletin, 35, 260–68.

Bazerman, M. H., Greene, J. D. (2010). In favor of clear thinking: Incorporating moral rules into a wise cost-benefit analysis. Perspectives on Psychological Science, 5(2), 209-212.

Bennis, W. M., Medin, D. L., Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. Perspectives on Psychological Science, 5(2), 187-202.

Białek, M., De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity. Judgment and Decision making, 12(2), 148-167.

Bicchieri, C. (2016). Norms in the wild: How to diagnose, measure, and change social norms. Oxford University Press.

Bina, F. (2022). Models of moral decision-making: Recent advances and normative relevance. Teoria, 42(2), 201-214.

Bina, F., Liberti M., Reichlin M., Songhorian S., Vaccarezza M. S. (unpublished manuscript). Individual moral progress: A virtue-based approach.

Birch, J. (2021). Toolmaking and the evolution of normative cognition. Biology & Philosophy, 36(1), 4.

Bird, D. W., Bird, R. B., Codding, B. F., Zeanah, D. W. (2019). Variability in the organization

and size of hunter-gatherer groups: Foragers do not live in small-scale societies. Journal of Human Evolution, 131, 96-108.

Bloom, P. (2010). How do morals change?. Nature, 464(7288), 490-490.

Boehm, C. (2012). Moral origins: The evolution of virtue, altruism, and shame. Basic Books.

Böhm, R., Rusch, H., Baron, J. (2020). The psychology of intergroup conflict: A review of theories and measures. Journal of Economic Behavior & Organization, 178, 947-962.

Bond, R., Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. Psychological Bulletin, 119(1), 111.

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. Journal of Cognitive Neuroscience, 18(5), 803–817.

Bornschier, V. (2002). Changing income inequality in the second half of the 20th century? Preliminary findings and propositions for explanations. Journal of World-Systems Research, 100-127.

Boudry, M., Vlerick, M., Edis, T. (2020). The end of science? On human cognitive limitations and how to overcome them. Biology & Philosophy, 35, 1-16.

Bowles, S., Gintis, H. (2013). A cooperative species. Princeton University Press.

Boyd, R., Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology, 13, 171–195.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. Journal of Social Issues, 55(3), 429-444.

Brink, D. O. (1989). Moral realism and the foundations of ethics. Cambridge University Press.

Brink, D. O. (2014). Principles and intuitions in ethics: Historical and contemporary perspectives. Ethics, 124(4), 665-694.

Broad, C. D. (1930). Five types of ethical theory. Routledge.

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. Nature, 425, 297–299.

Brosnan, S. F. (2006). Nonhuman species' reactions to inequity and their implications for fairness. Social Justice Research, 19, 153-185.

Buchanan, A. (2020). Our moral fate: Evolution and the escape from tribalism. MIT Press.

Buchanan, A., Powell, R. (2015). The limits of evolutionary explanations of morality and their implications for moral progress. Ethics, 126(1), 37-67.

Buchanan, A., Powell, R. (2016). Toward a naturalistic theory of moral progress. Ethics, *126*(4), 983-1014.

Buchanan, A., Powell, R. (2017). De-moralization as emancipation: Liberty, progress, and the evolution of invalid moral norms. Social Philosophy and Policy, 34(2), 108-135.

Buchanan, A., Powell, R. (2018). The evolution of moral progress: A biocultural theory. Oxford University Press.

Buller, D. J. (1998). Etiological theories of function: A geographical survey. Biology and Philosophy, 13, 505-527.

Buller, D. J. (2005). Adapting minds: Evolutionary psychology and the persistent quest for human nature. MIT Press.

Buss, D. M. (2019). Evolutionary psychology: The new science of the mind. Routledge.

Buss, D. M., Haselton, M. G., Shackelford, T. K., Bleske, A. L., Wakefield, J. C. (1998). Adaptations, exaptations, and spandrels. American Psychologist, 53(5), 533.

Campbell, R. (2017). Learning from moral inconsistency. Cognition, 167, 46-57.

Campbell, R., Kumar, V. (2012). Moral reasoning on the ground. Ethics, 122(2), 273-312.

Campbell, R., Kumar, V. (2013). Pragmatic naturalism and moral objectivity. Analysis, 73(3), 446-455.

Campbell, R., Woodrow, J. (2003). Why Moore's open question is open: The evolution of moral supervenience. The Journal of Value Inquiry, 37 (3), 353-372.

Carruthers, P. E., Laurence, S. E., Stich, S. E. (2005). The innate mind: Structure and contents. Oxford University Press.

Casal, S., Guala, F., Mittone, L. (2019). On the transparency of nudges: An experiment. *CEEL* working paper n. 1902.

Casebeer, W. D. (2003). Natural ethical facts: Evolution, connectionism, and moral cognition. MIT Press.

CDC (2020). Sexual assault awareness. Centers for Disease Control and Prevention, https://www.cdc.gov/injury/features/sexual-violence/index.html.

Cecchini, D. (2021). Dual-process reflective equilibrium: rethinking the interplay between intuition and reflection in moral reasoning. Philosophical Explorations, 24(3), 295-311.

Choi, H., Oishi, S. (2020). The psychology of residential mobility: A decade of progress. Current opinion in psychology, 32, 72-75.

Choi, J. K., Bowles, S. (2007). The coevolution of parochial altruism and war. Science, 318(5850), 636-640.

Christensen, M. B, Hallum, C., Maitland, A., Parrinello, Q., Putaturo, C. (2023). Survival of the richest. Oxfam International briefing paper, DOI: 10.21201/2023.621477.

Chappell, S.-G. (2022). Epiphanies. An ethics of experience. Oxford University Press.

Chomsky, N. (1975). Reflections on Language. Pantheon.

Cohen, M. A., Zenko, M. (2019). Clear and present safety. Yale University Press.

Colby, A., Damon, W. (1992). Some do care: Contemporary lives of moral commitment. Free Press.

Colburn, B. (2011). Autonomy and adaptive preferences. Utilitas, 23(1), 52-71.

Coleman, J. S. (1990). Foundations of social theory. Harvard University Press.

Conway P., Gawronski B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. Journal of Personality and Social Psychology, 104(2), 216.

Corr, P. J., Hargreaves Heap, S. P., Seger, C. R., Tsutsui, K. (2015). An experiment on individual 'parochial altruism'revealing no connection between individual 'altruism'and individual 'parochialism'. Frontiers in Psychology, 6, 1261.

Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (eds.), The adapted mind: Evolutionary psychology and the generation of culture. Oxford University Press. 163–228.

Cosmides, L., & Tooby, J. (2005). Neurocognitive adaptations designed for social exchange. In D. M. Buss (ed.), The Handbook of Evolutionary Psychology. Wiley, 584–627.

Cummins, D. D. (1996a). Evidence of deontic reasoning in 3- and 4-year-olds. Memory and Cognition, 24, 823–829.

Cummins, D. D. (1996b). Evidence for the innateness of deontic reasoning. Mind & Language, 11, 160–190.

Croce, M., Vaccarezza, M.S. (2017). Educating through exemplars: Alternative paths to virtue. Theory and Research in Education 15(1), 5-19.

Crockett, M. J. (2013). Models of morality. Trends in Cognitive Sciences, 17(8), 363-366.

Crosthwaite, J. (1995). Moral expertise: A problem in the professional ethics of professional ethicists. Bioethics, *9*(4), 361-379.

Crutchfield, P. (2021). Moral enhancement and the public good. Routledge.

Curry, O. S. (2016). Morality as cooperation: A problem-centred approach. In Shackelford. T. K., Hansen, D. (eds.), The evolution of morality. Springer, 25-51.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. Cognition, 108(2), 353–380.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. Personality and Social Psychology Review, 17(3), 273-292.

Cushman, F. (2020). Rationalization is rational. Behavioral and Brain Sciences, 43, e28.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. Psychological Science, 17(12), 1082-1089.

Cushman, F., Gray, K., Gaffey, A., Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. Emotion, 12(1), 2.

Cushman, F., Murray, D., Gordon-McKeon, S., Wharton, S., Greene, J. D. (2012). Judgment before principle: engagement of the frontoparietal control network in condemning harms of omission. Social Cognitive and Affective Neuroscience, 7(8), 888-895.

Cushman F., Morris A. 2015, Habitual control of goal selection in humans, *Proceedings of the National Academy of Sciences*, 112(45), 13817-13822.

Cushman, F., Kumar, V., Railton, P. (2017). Moral learning: Psychological and philosophical perspectives. Cognition, *167*, 1-10.

Damasio, A. R. (1994). Descartes' error. Putnam.

Darnell, C., Fowers, B.J., Kristjansson, K. (2022). A multifunction approach to assessing Aristotelian phronesis (practical wisdom). Personality and Individual Differences 196.

Darnell, C., Gulliford, L., Kristjánsson, K., Paris, P. (2019). Phronesis and the knowledge-action gap in moral psychology and moral education: A new synthesis?. Human Development 62(3), 1-29.

Dawkins, R. (1976). The selfish gene. Oxford University Press.

Dayan, P. (2012). How to set the switches on this thing. Current Opinion in Neurobiology, 22(6), 1068-1074.

De Caro, M., Macarthur, D. (eds.). (2010). Naturalism and normativity. Columbia University Press.

De Caro, M., Marraffa, M., Vaccarezza, M.S. (2021). The priority of phronesis. How to rescue virtue theory from its critics. In M. De Caro, M.S. Vaccarezza (eds.), Practical Wisdom. Philosophical and Psychological Perspectives. Routledge, 29-51.

De Dreu, C. K., Fariña, A., Gross, J., Romano, A. (2022). Prosociality as a foundation for intergroup conflict. Current Ppinion in Psychology, 44, 112-116.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. Thinking & Reasoning, 20(2), 169-187.

de Waal, F. (2006). Primates and philosophers: How morality evolved. Princeton University Press.

Deaton, A. (2013). The great escape: Health, wealth, and the origins of inequality. Princeton University Press.

Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. Psychological inquiry, 11(4), 227-268.

Dennett, D. C. (2003). *Freedom evolves*. Penguin.

Dickert, S., Västfjäll, D., Kleber, J., Slovic, P. (2012). Valuations of human lives: normative expectations and psychological mechanisms of (ir)rationality. Synthese, 189(1), 95-105.

Doris, J. M. (1998). Persons, situations, and virtue ethics. Noûs, 32, 504–530.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., Boakes, R. A., (1995). Motivational control after extended instrumental training. Animal Learning & Behavior, 23(2), 197-206.

Dolan, R. J., Dayan, P. (2013). Goals and habits in the brain. Neuron, 80(2), 312-325.

Doris, J. M. (2002). Lack of character: Personality and moral behavior. Cambridge University Press.

Douglas, T. (2008). Moral enhancement. Journal of Applied Philosophy, 25(3), 228-245.

Dreyfus, H. L., Dreyfus, S. E. (1991). Towards a phenomenology of ethical expertise. Human studies, 229-250.

Dunbar, R. I. (1993). Coevolution of neocortical size, group size and language in humans. Behavioral and brain sciences, 16(4), 681-694.

Dunbar, R. (2010). How many friends does one person need?. Harvard University Press.

Dworkin, R. (1984). Rights as Trumps. In Waldron J. (ed.). Theories of Rights. Oxford University Press, 153–67.

Dwyer, S., Huebner, B., Hauser, M. D. (2010). The linguistic analogy: Motivations, results, and speculations. Topics in Cognitive Science, 2(3), 486-510.

Easterbrook, G. (2018). It’s Better Than It Looks. Reasons for Optimism in an Age of Fear. PublicAffairs.

Elison, J. (2005). Shame and guilt: A hundred years of apples and oranges. New Ideas in Psychology, 23(1), 5-32.

Enos, R. D. (2017). The space between us: Social geography and politics. Cambridge University Press.

Ensminger, J., Henrich, J. (eds.). (2014). Experimenting with social norms: Fairness and punishment in cross-cultural perspective. Russell Sage Foundation.

Evans, J. (2017). A working definition of moral Progress. Ethical theory and moral practice, 20, 75-92.

Evans, J. S. B. (2019). Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning. Thinking & Reasoning, 25(4), 383-415.

Evans, J. S. B., Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing

the debate. Perspectives on Psychological Science, 8(3), 223-241.

Faulkner, J., Schaller, M., Park, J. H., Duncan, L. A. (2004). Evolved disease-avoidance mechanisms and contemporary xenophobic attitudes. Group Processes & Intergroup Relations, 7(4), 333-353.

Fehr, E., Fischbacher, U., Von Rosenbladt, B., Schupp, J., Wagner, G. G. (2002). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative survey. CEPR Discussion Papers 122 (141), 519-42.

Fessler, D. (2004). Shame in two cultures: Implications for evolutionary approaches. Journal of Cognition and Culture, 4(2), 207-262.

Festinger, L. (1957). A theory of cognitive dissonance. Stanford University Press.

Fine, C. (2006). Is the emotional dog wagging its rational tail, or chasing it? Reason in moral judgment. Philosophical Explorations, 9(1), 83-98.

Fitzgerald, D. K. (2008). Every farm a factory: The industrial ideal in American agriculture. Yale University Press.

FitzPatrick, W. J. (2015). Debunking evolutionary debunking of ethical realism. Philosophical Studies, 172, 883-904.

FitzPatrick, W. J. (2019). Moral progress for evolved rational creatures. Analyse & Kritik, 41(2), 217-238.

Flanagan, O. J. (1991). Varieties of moral personality: Ethics and psychological realism. Harvard University Press.

Fodor, J. A. (1983). The modularity of mind. MIT Press.

Frankena, W. (1967/1970). The concept of morality. In Wallace, G., Walker, A. D. (1970) (eds.), The definition of morality. Methuen, 146-173.

Fukuyama, F. (2002). Our posthuman future. Farrar, Straus and Giroux.

Fukuyama, F. D. (2012). The Origins of Political Order. Profile Books.

Gabennesch, H. (1990). The perception of social conventionality by children and adults. Child Development, 61, 2047–2059.

Gallagher, S. (2013). The socially extended mind. Cognitive Systems Research, 25, 4-12.

Garson, J. (2008). Function and teleology. In S. Sarkar & A. Plutynski (eds.). A companion to the philosophy of biology. Blackwell, 525–549.

Gaus, G. (2016). The tyranny of the ideal. Princeton University Press.

Gert, B., Gert, J. The Definition of Morality. In E. N. Zalta (ed.). The Stanford Encyclopedia of Philosophy [Online].

Gewirth, A. (1978). Reason and morality. University of Chicago Press.

Gewirth, A. (1996). The community of rights. University of Chicago Press.

Gigerenzer, G. (2007). Gut feelings: The intelligence of the unconscious. Penguin.

Gigerenzer, G. E., Hertwig, R. E., Pachur, T. E. (2011). Heuristics: The foundations of adaptive behavior. Oxford University Press.

Giubilini, A. (2022). Abortion, democracy, and erring on the side of freedom. Practical Ethics [Online].

Godfrey-Smith, P. (1994). A modern history theory of functions. Noûs, 28(3), 344-362.

Goldsmith, J. L., Posner, E. A. (2005). The limits of international law. Oxford University Press.

Goodwin, G. P. (2017). Is morality unified, and does this matter for moral reasoning?. In Bonnefon, J. F., Trémolière, B. (eds.). Moral inferences. Psychology Press, 17-44.

Gould, S. J., Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London B, 205, 581-98.

Gould, S. J., Vrba, E. S. (1982). Exaptation – A missing term in the science of form. Paleobiology, 8(1), 4-15.

Gowdy, J. (1999). Hunter-gatherers and the mythology of the market. In Lee, R. B., Daly, R. H., Daly, R. (eds.). The Cambridge encyclopedia of hunters and gatherers. Cambridge University Press, 391-398.

Granovetter, M. S. (1973). The strength of weak ties. American Journal of Sociology, 78(6), 1360-1380.

Gray, K., Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agent and moral patients. Journal of Personality and Social Psychology, 96(3), 505–520.

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. Annual Review of Neuroscience, 31, 359-387.

Greene, J. D. (2002). The terrible, horrible, no good, very bad truth about morality and what to do about it. Doctoral thesis. Princeton University.

Greene, J. (2003). From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology?. Nature Reviews Neuroscience, 4(10), 846-850.

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (ed.), Moral psychology: The neuroscience of morality: Emotion, disease, and development (Vol. 3). MIT Press, 35-79.

Greene, J. D. (2013). Moral tribes: Emotion, reason, and the gap between us and them. Penguin.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. Ethics, 124(4), 695-726.

Greene, J. D. (2015). The rise of moral cognition. Cognition, 135, 39-42.

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. Cognition, 167, 66-77.

Greene, J. D. (forthcoming). Dual-process moral judgment beyond fast and slow. Behavioral and Brain Sciences.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron, 44, 389–400.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition, 111(3), 364-371.

Greene, J.D., Young, L. (2020). The Cognitive Neuroscience of Moral Judgment and Decision-Making. In M.S. Gazzaniga (ed.). The Cognitive Neuroscience, Volume 6. MIT Press.

Griffiths, P. E., Machery, E., Linquist, S. (2009). The vernacular concept of innateness. Mind & Language, 24, 605–630.

Gürçay, B., Baron J. (2017). Challenges for the sequential two-system model of moral judgement. *Thinking & Reasoning*, 23(1), 49-80.

Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. Journal of Personality and Social Psychology, 104(2), 354.

Habermas, J. (1990). Moral consciousness and communicative action. MIT Press.

Hacker-Wright, J. (2015). Skill, practical wisdom, and ethical naturalism. Ethical Theory and Moral Practice, 18, 983-993.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review, 108(4), 814-834.

Haidt, J. (2012). The righteous mind: Why good people are divided by politics and religion. Pantheon Books.

Haidt, J., Koller, S. H., Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog?. Journal of Personality and Social Psychology, 65(4), 613.

Han, H. (2017). Neural correlates of moral sensitivity and moral judgment associated with brain circuitries of selfhood: A meta-analysis. Journal of Moral Education, 46(2), 97–113.

Hardin, G. J. (1977). The limits of altruism: An ecologist's view of survival. Bloomington.

Hare, R. M. (1952). The language of morals. Oxford University Press.

Hare, R. M. (1963). Freedom and reason. Oxford University Press.

Hare, R. M. (1972). The argument from received opinion. In Essays on philosophical method. University of California Press.

Hare, R. M. (1981). Moral thinking. Oxford University Press.

Harris, J. (2012). What it's like to be good. Cambridge Quarterly of Healthcare Ethics, 21(3), 293-305.

Harris, J. (2016). How to be good: The possibility of moral enhancement. Oxford University Press.

Harris, P. L., & Nuñez M. (1996). Understanding of permission rules by preschool children. Child Development, 67, 1572–1591.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. Proceedings of the Aristotelian Society, 99(3), 315–331.

Harris, J. (2012). What it's like to be good. Cambridge Quarterly of Healthcare Ethics, 21(3), 293-305.

Haselton, M. G., Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. Personality and Social Psychology Review, 10(1), 47-66.

Haslanger, S. (2015). Social structure, narrative and explanation. Canadian Journal of Philosophy, 45(1), 1-15.

Hauser, M. D. (2006). Moral minds: How nature designed our universal sense of right and wrong. Ecco.

Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. American Economic Review, 90(4), 973-979.

Henrich, J. (2015). The secret of our success. Princeton University Press.

Henrich, J. (2020). The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous. Penguin.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. American Economic Review, 91(2), 73-78.

Henrich, J. P., Boyd, R., Fehr, E., Bowles, S., Camerer, C., Gintis, H. (eds.). (2004). Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies. Oxford University Press.

Henrich, J., Heine, S. J., Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29-29.

Herman, B. (1993). The practice of moral judgment. Harvard University Press.

Hermann, J. (2019). The dynamics of moral progress. Ratio, 32(4), 300-311.

Hindriks, F., Sauer, H. (2020). The mark of the moral: Beyond the sentimentalist turn. Philosophical Psychology, 33(4), 569-591.

Hoffman, M. L. (2000). Empathy and moral development: Implications for caring and justice. Cambridge University Press.

Hogarth, R. M. (2001). Educating intuition. University of Chicago Press.

Hofstede, G. H. (2003). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations (2nd ed.). Sage.

Hopster, J. (2020). Explaining historical moral convergence: the empirical case against realist intuitionism. Philosophical Studies, 177(5), 1255-1273.

Huemer, M. (2016). A liberal realist answer to debunking skeptics: the empirical case for realism. Philosophical Studies, 173, 1983-2010.

Huemer, M. (2019). Dialogues on ethical vegetarianism. Routledge.

Hume, D. (1739/2000). A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects. A critical edition. (D. F. Norton. M. J. Norton, eds.). Oxford University Press.

Huxley, H. (1893/2009), Evolution and Ethics. Princeton University Press.

Inglehart, R. F. (2018). Cultural evolution: People's motivations are changing, and reshaping the world. Cambridge University Press.

Jamieson, D. (2002). Is there progress in morality?. *Utilitas*, 14(3), 318-338.

Johnson, N. D., Mislin, A. A. (2011). Trust games: A meta-analysis. Journal of economic psychology, 32(5), 865-889.

Johnson, M. (2020). Moral habit. In F. Caruana, I. Testa (eds.). Habits: Pragmatist approaches from cognitive science, neuroscience, and social theory. Cambridge University Press, 274-276.

Joyce, R. (2007). The evolution of morality. MIT Press.

Jost, J. T., Glaser, J., Kruglanski, A. W., Sulloway, F. J., (2003). Political conservatism as motivated social cognition. Psychological Bulletin, 3, 339-375.

Jubilee Centre for Character and Virtues (2022). The Jubilee Centre framework for character education in schools, Birmingham: University of Birmingham [Online]. Available at: https://www.jubileecentre.ac.uk/userfiles/jubileecentre/pdf/character-education/Framework%20for%20Character%20Education.pdf

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., Tracey, I. (2012). The neural

basis of intuitive and counterintuitive moral judgment. Social Cognitive and Affective Neuroscience, 7(4), 393-402.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. American Psychologist, 58(9), 697-720.

Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

Kauppinen, A. (2014). Ethics and Empirical Psychology–Critical Remarks to Empiricall Informed Ethics. In Christen, M. E., van Schaik, C. E., Fischer, J. E., Huppenbauer, M. E., & Tanner, C. E. (2014). Empirically informed ethics: Morality between facts and norms. Springer, 279-305.

Kelly, D., Stich, S., Haley, K. J., Eng, S. J., Fessler, D. M. (2007). Harm, affect, and the moral/conventional distinction. Mind & Language, 22(2), 117-131.

Kennett, J., Matthews, S. (2009). Mental timetravel, agency and responsibility. In M. Broome & L. Bortolotti (eds.), Psychiatry as cognitive neuroscience: Philosophical perspectives. Oxford University Press, 327–350.

Keramati, M., Dezfouli, A., Piray P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Computational Biology, 7(5), e1002055.

Kitcher, P. (1993). Function and design. Midwest Studies in Philosophy, 18, 379-397.

Kitcher, P. (2005). Biology and ethics. In D. Copp (ed.), The Oxford Handbook of Ethical Theory. Oxford University Press, 163-185.

Kitcher, P. (2014). The ethical project. Harvard University Press.

Kitcher P. (2015). Experimental animals. Philosophy and Public Affairs, 43(4), 287-311.

Kitcher, P. (2017). Social progress. Social Philosophy and Policy, 34(2), 46-65.

Kitcher, P. (2021). Moral progress. Oxford University Press.

Klenk, M., Sauer, H. (2021). Moral judgement and moral progress: The problem of cognitive control. Philosophical Psychology, 34(7), 938–961.

Knafo, A., Schwartz, S. H., Levine, R. V. (2009). Helping strangers is lower in embedded cultures. Journal of Cross-Cultural Psychology, 40(5), 875-879.

Kohlberg, L. (1981, 1984). Essays on moral development (Volumes I and II). San Francisco: Harper & Row.

Kool, W., Gershman, S. J., Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. Psychological Science, 28(9), 1321-1333.

Kool, W., Cushman, F. A., Gershman, S. J. (2018). Competition and cooperation between multiple reinforcement learning systems. In Morris, R. W., Bornstein, A., & Shenhav, A. (eds.). Goal-directed decision making: Computations and neural circuits. Academic Press,

153-178.

Koop, G. J. (2013). An assessment of the temporal dynamics of moral decisions. Judgment and Decision Making, 8(5), 527.

Korsgaard, C. M. (1996). The sources of normativity. Cambridge University Press.

Krebs, D. (2011). The origins of morality: An evolutionary account. Oxford University Press.

Kristjánsson, K. (2007). Aristotle, emotions, and education. Routledge.

Kristjánsson, K. (2015). Aristotelian character education. Routledge.

Kristjánsson, K. (2017). Awe: An Aristotelian analysis of a non-Aristotelian virtuous emotion. Philosophia 45(1), 125-142.

Kristjánsson, K. (2020). Flourishing as the aim of education: A neo-Aristotelian view. Routledge.

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodal. Perspectives on Psychological Science, 8(3), 242-247.

Kumar, V. (2017). Moral vindications. Cognition, 167, 124-134.

Kumar, V., Campbell, R. (2022). A better ape: The evolution of the moral mind and how it made us human, Oxford University Press.

Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480-498.

Kurth, C. (2016). Anxiety, normative uncertainty, and social regulation. *Biology & Philosophy*, *31*, 1-21.

Kurzban, R., Leary, M. R. (2001). Evolutionary origins of stigmatization: The functions of social exclusion. Psychological Bulletin, 127(2), 187.

Lapsley, D. (2016). On the prospects for Aristotelian character education. Journal of Moral Education, 45(4), 502-515.

Lee, W. E. (2016). Waging war: Conflict, culture, and innovation in world history. Oxford University Press.

Li, N. P., van Vugt, M., Colarelli, S. M. (2018). The evolutionary mismatch hypothesis: Implications for psychological science. Current Directions in Psychological Science, 27(1), 38-44.

Lickliter, R., Honeycutt, H. (2003). Developmental dynamics: toward a biologically plausible evolutionary psychology. Psychological Bulletin, 129(6), 819.

Luco, A. (2014). The definition of morality: Threading the needle. Social Theory and Practice, 361-387.

Luco, A. (2019). How moral facts cause moral progress. Journal of the American Philosophical Association, 5(4), 429-448.

Lun, J., Oishi, S., Tenney, E. R. (2012). Residential mobility moderates preferences for egalitarian versus loyal helpers. Journal of Experimental Social Psychology, 48(1), 291-297.

Ma, V., Schoeneman, T. J. (1997). Individualism versus collectivism: A comparison of Kenyan and American self-concepts. Basic and applied social psychology, 19(2), 261-273.

Machery, E., Mallon, R. (2010). Evolution of morality. Doris, J. M., & Moral Psychology Research Group. The Moral Psychology Handbook. Oxford University Press.

Machery, E., Stich, E. (2022). The Moral/Conventional Distinction. In E. N Zalta (ed.). The Stanford Encyclopedia of Philosophy [Online].

Mackie, J. L. (1977). Ethics: Inventing right and wrong. Penguin.

Macklin, R. (1977). Moral progress. Ethics, 87(4), 370-382.

Madva, A. (2016). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. Ergo, 3, 27, 701-728.

Mallon, R., Weinberg, J. (2006). Innateness as closed-process invariantism. Philosophy of Science, 73, 323–344.

Marlowe, F. W., Berbesque, J. C., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Ensminger, G., Gurven, M., Gwako, E., Henrich, J., Henrich, N., Lesogorol, C., McElreath, R., Tracer, D. (2008). More 'altruistic' punishment in larger societies. Proceedings of the Royal Society B: Biological Sciences, 275(1634), 587-592.

May, J. (2013). Because I believe it's the right thing to do. Ethical Theory and Moral Practice, 16, 791-808.

McCloskey, D. N. (2010). The bourgeois virtues: Ethics for an age of commerce. University of Chicago Press.

McGrath, S. (2008). Moral disagreement and moral expertise. In Shafer-Landau, R. (ed.). Oxford Studies in Metaethics: Volume 3. Oxford University Press, 87-108.

McDonald, M. M., Navarrete, C. D., Van Vugt, M. (2012). Evolution and the psychology of intergroup conflict: The male warrior hypothesis. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1589), 670-679.

McNamara, R. A., Willard, A. K., Norenzayan, A., Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. Cognition, 182, 95-108.

Mercier, H., Sperber, D. (2017). The enigma of reason. Harvard University Press.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. Trends in Cognitive Sciences, 11(4), 143-152.

Miller, C. B. (2014). Character and moral psychology. Oxford University Press.

Miller, C. B. (2017). The character gap: How good are we?. Oxford University Press.

Miller, C. B. (2021). Flirting with skepticism about practical wisdom. In M. De Caro, M.S. Vaccarezza (eds.), Practical Wisdom. Philosophical and psychological perspectives. Routledge, 52-69.

Millikan, R. G. (1984). Language, thought and other biological categories. MIT Press.

Millikan, R. G. (1989). In defense of proper functions. Philosophy of Science, 56(2), 288-302.

Moody-Adams, M. M. (2017). Moral progress and human agency. Ethical Theory and Moral Practice, 20(1), 153-168.

Moody-Adams, M. M. (1999). The idea of moral progress. Metaphilosophy, 30(3), 168-185.

Moore, G. E. (1903/1993). Principia Ethica. Cambridge University Press.

Musschenga, A. W., Meynen, G. (2017). Moral progress: An introduction. Ethical Theory and Moral Practice, 20, 3-15.

Muthukrishna, M., Henrich, J., Slingerland, E. (2021). Psychology as a historical science. Annual Review of Psychology, 72, 717-749.

Nagel, T. (2012). Mind and cosmos: Why the materialist neo-Darwinian conception of nature is almost certainly false. Oxford University Press.

Navarrete, C. D., Fessler, D. M. (2006). Disease avoidance and ethnocentrism: The effects of disease vulnerability and disgust sensitivity on intergroup attitudes. Evolution and Human Behavior, 27(4), 270-282.

Neander, K. (1991). The teleological notion of 'function'. Australasian Journal of Philosophy, 69(4), 454-468.

Neuberg, S. L., Schaller, M. (2016). An evolutionary threat-management approach to prejudices. Current Opinion in Psychology, *7*, 1-5.

Neurath, O. (1921). Anti-Spengler. Reprinted in Neurath, M., Cohen, R. S. (eds.). (1973). Empiricism and sociology. Reidel.

Nichols, S. (2004). Sentimental rules: On the natural foundations of moral judgment. Oxford University Press.

Norberg, J. (2020). Open: The story of human progress. Atlantic Books.

Norenzayan, A., Shariff, A. F., Gervais, W. M., Willard, A. K., McNamara, R. A., Slingerland, E., Henrich, J. (2016). The cultural evolution of prosocial religions. Behavioral and brain sciences, 39, e1.

Nowell-Smith, P. (1957). Ethics. Philosophical Library.

North, D. C., Wallis, D. J., Weingast, B. R. (2009). Violence and Social Orders. Cambridge

University Press.

Nozick, R. (1974). Anarchy, state, and utopia. Basic Books.

Nussbaum, M. C. (2000). Women and human development: The capabilities approach. Cambridge University Press.

Nussbaum, M. C. (2011). Creating capabilities. Harvard University Press.

Oaten, M., Stevenson, R. J., Case, T. I. (2011). Disease avoidance as a functional basis for stigmatization. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1583), 3433-3452.

Page, L. (2022). *Optimally Irrational: The Good Reasons We Behave the Way We Do*. Cambridge University Press.

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. Journal of Cognitive Neuroscience, 23(10), 3162-3180.

Pascual, L., Gallardo-Pujol, D., Rodrigues, P. (2013). How does morality work in the brain? A functional and structural perspective of moral behavior. Frontiers in Integrative Neuroscience, 7(65), 1–8.

Patil, I., Zucchelli, M. M., Kool, W., Campbell, S., Fornasier, F., Calò, M., Cikara, M., Cushman, F. (2021). Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. Journal of Personality and Social Psychology, 120(2), 443-460.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. Cognitive science, 36(1), 163-177.

Persson, I., & Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. Journal of Applied Philosophy, 25(3), 162-177.

Persson, I., Savulescu, J. (2012). Unfit for the future: The need for moral enhancement. Oxford University Press.

Persson, I., Savulescu, J. (2017). Moral hard-wiring and moral enhancement. Bioethics, 31(4), 286-295.

Phillips, J., Cushman, F. (2017). Morality constrains the default representation of what is possible. Proceedings of the National Academy of Sciences, 114(18), 4649-4654.

Pietraszewski, D, Wertz, A. E. (2022). Why Evolutionary Psychology Should Abandon Modularity. Perspectives on Psychological Science, 17(2), 465-490.

Piketty, T., Saez, E. (2003). Income inequality in the United States, 1913–1998. The Quarterly

journal of economics, 118(1), 1-41.
Piketty, T., Saez, E. (2014). Inequality in the long run. Science, 344(6186), 838-843.
Pinker, S. (1997). How the mind works. Norton.
Pinker, S. (2002). The blank slate: The modern denial of human nature. Penguin.
Pinker, S. (2011). The better angels of our nature: The decline of violence in history and its causes. Penguin.
Pinker, S. (2018). Enlightenment now: The case for reason, science, humanism, and progress. Penguin.
Pisor A. C., Surbeck M. (2019). The evolution of intergroup tolerance in nonhuman primates and humans. Evolutionary Anthropology 28(4), 210-223.
Powell, R., Buchanan, A. (2016). The Evolution of Moral Enhancement. In S. Clarke, J. Savulescu (eds.). The ethics of human enhancement: Understanding the debate. Oxford University Press, 239–260.
Prichard, H. A. (1949). Moral obligation. Oxford University Press.
Prinz, J. (2006). The emotional basis of moral judgments. Philosophical Explorations, 9(1), 29-43.
Prinz, J. (2006). Is the mind really modular?. Contemporary Debates in Cognitive Science, 14, 22-36.
Prinz, J. (2007). The emotional construction of morals. Oxford University Press.
Prinz, J. (2008). Is morality innate?. In Sinnott-Armstrong, W. E. (ed). Moral psychology, Vol 1. The evolution of morality: Adaptations and innateness. MIT Press, 367-406.
Prinz, J. J. (2012). Beyond human nature: How culture and experience shape the human mind. Norton.
Rai, T. S., Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. Psychological review, 118(1), 57.
Railton, P. (1986). Moral realism. The Philosophical Review, 95(2), 163-207.
Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. Ethics, 124(4), 813-859.
Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. Cognition, 167, 172-190.
Rawls, J. (1951). Outline of a decision procedure for ethics. The Philosophical Review, 60(2), 177-197.
Rawls, J. (1971). A Theory of Justice. Harvard University Press.
Rawls, J. (1980). Kantian constructivism in moral theory. The Journal of Philosophy, 77(9),

515-572.

Ross, W. D. (1930/2002). The right and the good. Oxford University Press.

Raz, J. (1986). The morality of freedom. Oxford University Press.

Rebonato, R. (2012). Taking liberties: A critical examination of libertarian paternalism. Springer.

Reger, J., Lind, M. I., Robinson, M. R., Beckerman, A. P. (2018). Predation drives local adaptation of phenotypic plasticity. Nature Ecology & Evolution, 2(1), 100-107.

Rehren, P., Sauer, H. (2022). Another brick in the wall? Moral education, social learning, and moral progress. Ethical Theory and Moral Practice, 1-16.

Reich, R. (2002). Bridging liberalism and multiculturalism in American education. University of Chicago Press.

Reichlin, M. (2018). On the idea of a 'method' in moral philosophy. Phenomenology and mind, (15), 60-69.

Reichlin, M. (2019). The moral agency argument against moral bioenhancement. Topoi, 38(1), 53-62.

Rhoads, S. A., Gunter, D., Ryan, R. M., Marsh, A. A. (2021). Global variation in subjective well-being predicts seven forms of altruism. Psychological Science, 32(8), 1247-1261.

Richerson, P. J., Boyd, R. (1999). Complex societies: The evolutionary origins of a crude superorganism. Human nature, 10, 253-289.

Ritchie, H., Rosado, P., Roser, M. (2017). Meat and dairy production. Our World in Data.

Rønnow-Rasmussen, T. (2017). On locating value in making moral progress. Ethical Theory and Moral Practice, 20, 137-152.

Rorty R. (1999). Human rights, rationality and sentimentality. In O. Savic (ed.). The politics of human rights. Verso, 67-83.

Rosling, H. (2018). Factfulness. Sceptre.

Royzman, E. B., Baron, J. (2002). The preference for indirect harm. Social Justice Research, 15, 165-184.

Runge, C. F. (1984). Institutions and the free rider: The assurance problem in collective action. The Journal of Politics, 46(1), 154-181.

Rustagi, D., Engel, S., Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. Science, *330*(6006), 961-965.

Samuels, R. (1998). Evolutionary psychology and the massive modularity hypothesis. British Journal for the Philosophy of Science, 49, 575–602.

Samuels, R. (2002). Nativism in cognitive science. Mind & Language, 17, 233–265.

Sanderse, W. (2015). An Aristotelian model of moral development. Journal of Philosophy of Education 49(3), 382-398.

Santos, H. C., Varnum, M. E., & Grossmann, I. (2017). Global increases in individualism. Psychological Science, 28(9), 1228-1239.

Sauer, H. (2012). Educated intuitions. Automaticity and rationality in moral judgement. Philosophical Explorations, 15(3), 255-275.

Sauer, H. (2017). Moral judgments as educated intuitions. MIT Press.

Sauer, H. (2019). Butchering benevolence moral progress beyond the expanding circle. Ethical Theory and Moral Practice, 22(1), 153–167.

Sauer, H. (2023). Moral teleology: A theory of progress. Routledge.

Sauer, H., Blunden, C., Eriksen, C., Rehren, P. (2021). Moral progress: Recent developments. Philosophy Compass, 16(10), e12769.

Saunders, L. F. (2016). Reason and emotion, not reason or emotion in moral judgment. Philosophical Explorations, 19(3), 252-267.

Scanlon, T. M. (1998). What we owe to each other. Harvard University Press.

Scanlon, T. (2018). Why does inequality matter?. Oxford University Press.

Schaefer, G. O. (2015). Direct vs. indirect moral enhancement. Kennedy Institute of Ethics Journal, 25(3), 261-289.

Schaefer, G. O., Savulescu, J. (2019). Procedural moral enhancement. Neuroethics, 12(1), 73–84.

Schinkel, A., de Ruyter D. J. (2017). Individual moral development and moral progress. Ethical Theory and Moral Practice, 20(1), 121–136.

Schulz, A. W. (2020). Enhancing thoughts: Culture, technology, and the evolution of human cognitive uniqueness. Mind & Language, 37(3), 465-484.

Schulz, J. F. (2022). Kin networks and institutional development. The Economic Journal, 132(647), 2578-2613.

Schulz, J. F., Bahrami-Rad, D., Beauchamp, J. P., Henrich, J. (2019). The Church, intensive kinship, and global psychological variation. Science, 366(6466), eaau5141.

Schwitzgebel, E., Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. Cognition, 141, 127-137.

Schwitzgebel, E., Rust, J. (2016). The behavior of ethicists. In Sytsma, J., Buckwalter, W. (eds.) (2016). A companion to experimental philosophy, Wiley, 225-233.

Segovia-Cuéllar, A., Del Savio, L. (2021). On the use of evolutionary mismatch theories in debating human prosociality. Medicine, Health Care and Philosophy, 24(3), 305-314.

Sen, A. (1985). Well-being, agency and freedom: The Dewey Lectures 1984. Journal of Philosophy, 82(4), 169-221.

Sen, A. (1999). Development as freedom. Knopf.

Sen, A. (2009). The idea of justice. Harvard University Press.

Severini, E. (2021). Moral progress and evolution: knowledge versus understanding. Ethical Theory and Moral Practice, 24(1), 87-105.

Shafer-Landau, R. (2003). Moral realism: A defence. Oxford University Press.

Shafer-Landau, R. (2012). Evolutionary debunking, moral realism and moral knowledge. Journal of Ethics and Social Philosophy, 7, 1, 1-37.

Shariff, A. F., Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. Psychological science, 18(9), 803-809.

Shariff, A. F., Norenzayan, A. (2011). Mean gods make good people: Different views of God predict cheating behavior. The International Journal for the Psychology of Religion, 21(2), 85-96.

Shenhav, A., Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron, 67(4), 667–677.

Shweder, R. A., Mahapatra, M., Miller, J. (1987). Culture and moral development. In J. Kagan, S. Lamb (eds.), The emergence of morality in young children. University of Chicago Press, 1–82.

Sidgwick, H. (1907/1981). The methods of ethics. Hackett.

Sinclair, N. (2012). Metaethics, teleosemantics and the function of moral judgments. Biology and Philosophy, 27(5), 639–662.

Singer, P. (1972). Moral experts. Analysis, 32, 1155-117.

Singer, P. (1981/2011). The expanding circle: Ethics, evolution, and moral progress. Princeton University Press.

Singer, P. (2005). Ethics and intuitions. The Journal of Ethics, 9(3/4), 331-352.

Singer, P., Wells, D. (1984). The reproduction revolution: New ways of making babies. Oxford University Press.

Sinnott-Armstrong, W., Wheatley, T. (2014). Are moral judgments unified?. Philosophical Psychology, 27(4), 451-474.

Slote, M. (2007). The ethics of care and empathy. Routledge.

Smaldino, P. E., Lukaszewski, A., von Rueden, C., Gurven, M. (2019). Niche diversity can

explain cross-cultural differences in personality structure. Nature Human Behaviour, 3(12), 1276-1283.

Smetana, J. (1981). Preschool children's conceptions of moral and social rules. Child Development, 52, 1333–1336.

Smith, A. (1759/2004). The theory of moral sentiments. Cambridge University Press.

Smyth, N. (2017). The function of morality. Philosophical Studies, 174(5), 1127-1144.

Smyth, N. (2020). A genealogy of emancipatory values. *Inquiry*, 1-30.

Snow, N. E, Cole Wright, J., Warren, M. T. (2021). *Phronēsis* and whole trait theory: an integration. In M. De Caro, M.S. Vaccarezza (eds.), Practical Wisdom. Philosophical and Psychological Perspectives. Routledge.

Songhorian, S. (2019). The methods of neuroethics: Is the neuroscience of ethics really a new challenge to moral philosophy?. Rivista internazionale di Filosofia e Psicologia, 10(1), 1-15.

Songhorian, S., Guma, F., Bina, F., Reichlin, M. (2022). Moral progress: *Just* a matter of behavior?, Teoria, 42(2), 175-187.

Sperber, D. (1996). Explaining culture: A naturalistic approach. Blackwell.

Stanford, P. K. (2018). The difference between ice cream and Nazis: Moral externalization and the evolution of human cooperation. Behavioral and Brain Sciences, 41, 1-49.

Sterelny, K. (2010). Moral nativism: A sceptical response. Mind & Language, 25(3), 279-297.

Sterelny, K. (2012). *The evolved apprentice*. MIT press.

Sterelny, K. (2019). Evolutionary foundations for a theory of moral progress?. Analyse & Kritik, 41(2), 205-216.

Steutel, J., Carr, D. (1999). Virtue ethics and the virtue approach to moral education. In D. Carr, J. Steutel (eds.), Virtue ethics and moral education. Routledge, 3-18.

Stichter, M. (2007). Ethical expertise: The skill model of virtue. Ethical Theory and Moral Practice, 10, 183–194.

Stichter, M. (2018). The skillfulness of virtue: Improving our moral and epistemic lives. Cambridge University Press.

Stoks, R., Govaert, L., Pauwels, K., Jansen, B., De Meester, L. (2016). Resurrecting complexity: The interplay of plasticity and rapid evolution in the multiple trait response to strong changes in predation pressure in the water flea *Daphnia magna*. Ecology Letters, 19(2), 180-190.

Stotz, K. (2014). Extended evolutionary psychology: the importance of transgenerational developmental plasticity. Frontiers in Psychology, 5, 908.

Street, S. (2006). A Darwinian dilemma for realist theories of value. Philosophical Studies, 127, 109-166.

Sunstein, C. R. (2015). The ethics of nudging. Yale Journal on Regulation, 32, 413-450.

Swanton, C. (2016). Developmental virtue ethics. In J. Annas, D. Narvaez, N.E. Snow (eds.), Developing the virtues: Integrating perspectives. Oxford University Press, 116-134.

Symons D. (1992). On the use and misuse of Darwinism in the study of human behavior. In Barkow J., Cosmides L., Tooby J. (eds.), The adapted mind: Evolutionary psychology and the generation of culture. Oxford University Press, 137–159.

Tam, A. (2020). Why moral reasoning is insufficient for moral progress. Journal of Political Philosophy, 28(1), 73-96.

Thaler, R. H., Sunstein, C. R. (2008). Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Thompson, V. A., Turner, J. A. P., Pennycook, G. (2011). Intuition, reason, and metacognition. Cognitive psychology, 63(3), 107-140.

Tomasello, M. (2016). A natural history of human morality. Harvard University Press.

Tomasello, M., Vaish, A. (2013). Origins of human cooperation and morality. Annual Review of Psychology, 64, 231-255.

Tooby, J. (2020). Evolutionary psychology as the crystalizing core of a unified modern social science. Evolutionary Behavioral Sciences, 14(4), 390-403.

Tooby, J., Cosmides, L. (1990). On the universality of human nature and the uniqueness of the individual: The role of genetics and adaptation. Journal of Personality, *58*(1), 17-67.

Trémolière, B., Bonnefon, J. F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. Personality and Social Psychology Bulletin, 40(7), 923-930.

Triandis, H. C. (1995). Individualism and collectivism. Westview Press.

Tropp, L. (ed.). (2012). The Oxford handbook of intergroup conflict. Oxford University Press.

Turiel, E. (1983). The Development of Social Knowledge. Cambridge: Cambridge University Press.

Uchiyama, R., Spicer, R., Muthukrishna, M. (2022). Cultural evolution of genetic heritability. Behavioral and Brain Sciences, 45, e152.

Wallbott, H. G., Scherer, K. R. (1995). Cultural determinants in experiencing shame and guilt. Guilford Press.

Watkins, A. (2021). Testing for phenotypic plasticity. Philosophy, Theory, and Practice in Biology, 13, 3.

Waytz, A., Gray, K., Epley, N., Wegner, D. M. (2010). Causes and consequences of mind perception. Trends in Cognitive Sciences, 14(8), 383–388.

Welzel, C. (2007). Are levels of democracy affected by mass attitudes? Testing attainment and sustainment effects on democracy. International Political Science Review, 28(4), 397-424.

Welzel, C. (2013). Freedom rising. Cambridge University Press.

Welzel, C., Inglehart, R. (2010). Agency, values, and well-being: A human development model. Social Indicators Research, 97, 43-63.

Wilson, E.O. (1975). Sociobiology: The new synthesis. Harvard University Press.

Wrangham, R. W., Peterson, D. (1996). Demonic males: Apes and the origins of human violence. Houghton Mifflin Harcourt.

Wright, J. C., Warren, M. T., Snow, N. E. (2020). Understanding virtue: Theory and measurement. Oxford University Press.

Wright, L. (1976). Teleological explanations: An etiological analysis of goals and functions. University of California Press.

Wood, A. W. (1999). Kant's ethical thought. Cambridge University Press.

Wrangham, R. (2009). *Catching fire: how cooking made us human*. Basic Books.

Yamagishi, T., Mifune, N. (2016). Parochial altruism: Does it explain modern human group psychology?. Current Opinion in Psychology, 7, 39-43.

Ypi, L. (unpublished manuscript). The moral ought in 'as if' history.

Young, L., Dungan, J. (2012). Where in the brain is morality? Everywhere and maybe nowhere. Social neuroscience, 7(1), 1-10.

Zagzebski, L. T. (2017). Exemplarist moral theory. Oxford University Press.

Zmigrod, L. (2022). A psychology of ideology: Unpacking the psychological structure of ideological thinking. Perspectives on Psychological Science, 17(4), 1072-1092.

Zmigrod, L., Eisenberg, I. W., Bissett, P. G., Robbins, T. W., Poldrack, R. A. (2021). The cognitive and perceptual correlates of ideological attitudes: a data-driven approach. Philosophical Transactions of the Royal Society B, 376(1822), 20200424.