Ethics, Law and AI | 1

Roberto Redaelli [Ed.]

Moral Normativity in an Interdisciplinary Perspective

Humans, Animals & Artificial Intelligence

VERLAG KARL ALBER

Ethics, Law and AI

Edited by

Carmine Di Martino (Università degli Studi di Milano)
Federico L.G. Faroldi (Università di Pavia)
Roberto Redaelli (Università degli Studi di Milano)

Volume 1

Roberto Redaelli [Ed.]

Moral Normativity in an Interdisciplinary Perspective

Humans, Animals & Artificial Intelligence

VERLAG KARL ALBER

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de

ISBN 978-3-495-99428-3 (Print)
978-3-495-99429-0 (ePDF)

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN 978-3-495-99428-3 (Print)
978-3-495-99429-0 (ePDF)

Library of Congress Cataloging-in-Publication Data
Redaelli, Roberto
Moral Normativity in an Interdisciplinary Perspective
Humans, Animals & Artificial Intelligence
Roberto Redaelli (Ed.)
141 pp.
Includes bibliographic references.

ISBN 978-3-495-99428-3 (Print)
978-3-495-99429-0 (ePDF)

Online Version
Nomos eLibrary

1st Edition 2023

Table of Contents

Roberto Redaelli

Moral Normativity in Humans, Animals & Artificial Intelligence. An Introduction

Nowadays many disciplines are devoting particular attention from a variety of perspectives to the normative nature of our forms of life. From linguistics to jurisprudence, from anthropology to philosophy, from economics to neuroscience, the subject of moral normativity constitutes a Gordian knot of the present age, towards which the efforts of scientific and philosophical understanding are directed.

The following volume is addressed to achieving a better understanding of moral normativity. It collects works by primatologists, sociologists and philosophers of law, ethicists and phenomenologists, and their contribution to resolving issues regarding the normative profile of ethical concepts, judgments and reasons, i.e., the source of their binding force that guides the behaviour of the human agent.

A primary aspect of this issue, which we address in this volume, is the link between humans and animals, examined in a twofold direction of inquiry. On the one hand, the evolutionary perspective, which questions the natural history of our species, makes a decisive contribution to understanding the origin and nature of human moral normativity. On the other hand, using an ontogenetic perspective, it is possible to recognise forms of proto-normativity in children and animals.

The first two writings that open the volume refer to this framework of reflections. Richard Wrangham's valuable contribution *The transition from might to right: male-male conflict and the evolution of Homo duplex* is aimed at reconstructing the origins of moral normativity starting from an evolutionary explanation. In other words, it is a question of understanding why natural selection favors the extreme form of cooperation that morality represents. The hypothesis advanced by Wrangham is that adult males cooperated in executing tyrants using pre-conceived plans. Those alliances then became social means used to impose group norms that benefited those males. Following

this dynamic, Wrangham convincingly shows the emergence of the normative force of moral norms and together with it of the figure of the so-called *Homo duplex*, who is governed by the opposited motivations of selfishness and groupishness.

This first contribution is followed by that of Edoardo Fittipaldi: *Norms, Rights, Obligations: An Attempt at Empirical Reduction*. In this contribution, Fittipaldi conceptualizes norms in terms of psychical dispositions to experience *normative* emotions toward certain behaviors. In turn, *normative* emotions are conceptualized as emotions that emerge during primary socialization by virtue of the manner in which the child conceives of their caregiver, namely, much as monotheisms conceive of the One God. By using this hypothesis (which was first formulated by Bovet, Freud and Piaget), Fittipaldi reconceptualizes the notions of norm, right and obligation, and argues that it makes sense to speak of the existence of proto-norms and proto-rights, as well as of proto- and para-normative emotions, and so both in human and non-human animals.

A second direction of investigation is that developed by the contributions of John Drummond and Alessio Rotundo. These authors address the normative dimension of human being by making use of the phenomenological perspective. The masterful analyses carried out by John Drummond in *The Normativity of Norms* aim to highlight how the normativity of norms is rooted in the structures and goods inherent in the structures of rational experience, where chief among these structures, as Husserl affirmed, is intentionality, the mind's directedness to the world, and chief among these goods is what the author defines as truthfulness. In order to demonstrate this thesis, Drummond's reflections are based on a double distinction, that epistemic and practical norms, and with regard to practical norms, between first- and second-order norms.

Rotundo's contribution aims to highlight and discuss the positions of Maurice Merleau-Ponty and Raymond Aron on interpreting cultural values. Both authors converge around a critique of the relativistic viewpoint advanced by the anthropology of Claude Lévi-Strauss and propose the idea of an ethics predicated on plurality that is not at odds with a reflection on the meaning of rational humanity in history. As clearly shown by Rotundo this ethics assumes in Aron the form of a "pluralist anthropology" and in Merleau-Ponty of a theory of "historical symbolism". The question of moral normativity is thus

linked to the ability to hold together the plurality of different cultures without however falling into a form of relativism.

The last two writings that close the volume address the relationship between moral normativity and artificial intelligence. Roberto Redaelli's article *Artificial Intelligence and Quasi-Normativity* aims to investigate the notion of digital normativity, understood as the binding force exerted on the human subject by the predictions and standards established by artificial intelligent systems. In order to explain the AI binding force, Redaelli introduces the notion of quasi-normativity using the post-phenomenological perspective of Don Ihde, who defines AI in terms of quasi-other. With the notion of quasi-normativity the author intends to show how the models generated by AI already possess an injunctive force, linked to the predictive efficiency of algorithms, that redefines to a certain degree our space of freedom and directs our action.

The volume concludes with a contribution by Federico Faroldi who programmatically outlines a solution to the AI alignment problem starting from the Human-Compatible Approach developed by Stuart Russell. This promising approach proposes that intelligent agents be not required to maximize a simple given reward function attached to single aims, but to maximize the realization of human preferences, which are essentially uncertain. By developing this research direction, Faroldi intends to propose methods that will enable AI systems to comply not only with precise rules, but also with ethical principles and moral values. In this sense, the normative issue is addressed starting from a perspective that aims to account for the alignment of artificial intelligence with human values.

By approaching the normative issue from an interdisciplinary perspective, which includes both living beings and non-living things, the book aims to provide the reader with an overview of a series of problems that are increasingly urgent today, and due to which not only the present but also the future of our humanity is at stake.

Richard Wrangham

The transition from might to right: male-male conflict and the evolution of *Homo duplex*[1]

Charles Darwin's 1871 publication of ›*Sexual Selection and the Descent of Man*‹ launched new approaches to normative questions about morality. The concept of a sense of right and wrong had traditionally been explained by reference to divine powers, but evolutionary theory opened the way to investigations of morality using biological and social sciences. The problem was difficult, however. Natural selection was supposed to favor traits that benefit individuals and their kin, not non-kin. Morally right behavior, by contrast, often involves agents sacrificing their own immediate interests for the benefit of a larger group, many or most of whom are not genetically related to the agent. So a critical question became: what could explain the evolution of those moral tendencies underlying unselfish behavior that benefits non-kin? Darwin did not have an answer that satisfied him, and the question has continued to puzzle scholars to the present day.

The problem applies only to humans because although non-human animals can exhibit forms of morality including empathy, prosociality and mutualistic cooperation, only humans have moral tendencies that promote group benefits (de Waal, 2006; Engelmann et al., 2017; McAuliffe and Santos, 2018). Only humans, therefore, conform to Durkheim's (1973) notion of *Homo duplex* – a species in which individuals are torn between their motivations for selfishness on the one hand, and for promoting the interests of their social groups, on the other (Kluver et al., 2014). In the remainder of this essay I use »morality« to refer to the uniquely human forms of moral behavior that are self-sacrificial on behalf of a group or towards individual group members.

---

[1] This paper is an extended and revised version of my text *The Execution Hypothesis for the Evolution of a Morality of Fairness* published in: *Ethics & Politics* (2021) XXIII/2, pp. 261–282.

Several reasons suggest that humans have innate tendencies to behave morally. Emotions are clearly important in moral judgment, given that agents can be committed to moral decisions for which they are unable to produce any rational explanation (i.e. moral dumb-founding, Haidt, 2012). Furthermore in studies of moral decision-making, neural regions have been identified that are more engaged in the production of quick and automatic emotional responses than in slower, consciously reasoned reactions (Greene and Young, 2020). Such emotions are theorized to have an innate component in the form of a norm psychology, i.e. a tendency to acquire norms, comply with norms, and punish norm violators (Chudek and Henrich, 2011; Sripada and Stich, 2006).

Here I present a hypothesis that purports to explain the evolution of a norm psychology and the resulting moral behavior. The proposal, which I call the execution hypothesis, was initiated by Boehm (2008, 2012, 2018) and developed by Wrangham (2019a, 2021). Like many other theoretical investigations of the origins of morality it focuses on cooperation as the critical problem: why did natural selection favor the extreme form of cooperation that morality represents, such that individuals become willing to compromise their own immediate interests for the sake of a larger group?

1. Repression of competition: a critical condition for the evolution of cooperation.

The execution hypothesis can be traced to a theoretical analysis by the biologist Richard Alexander (1987). Alexander argued from first principles that cooperation can evolve among individuals only in the rare condition when a mechanism suppresses competition among them. His argument helped spawn the idea that the repression of competition is a necessary precondition for the evolution of cooperation in general, whether among individual organisms or sub-organismic units (Frank, 2003; Frank, 2013). Such units below the level of the individual include genes (cooperating as genomes) and cells (cooperating as metazoan organisms). The theoretical concept that solves the problem is called the enforcement of cooperation (Ågren et al., 2019; West et al., 2021). A vital question in all such cases is to explain why a mechanism occurs that represses internal

competition »in the face of the ubiquitous drive toward individual selfishness« (Frank, 2003, 694).

The most prominent and impactful form of individual selfishness found among the primates to which humans are related is competition for dominance among adult males. In Old World monkeys and apes male-male competition has led to males evolving multiple adaptations for fighting, including large bodies, strong muscles, long canine teeth and high motivation to respond to challenges with aggression. These adaptations have been favored because males who win fights and achieve high dominance status tend to achieve high evolutionary success in the form of elevated levels of paternity and offspring survival. Among humans' primate relatives, this competitive dynamic has invariably led to one male dominating all other males in his social group. The male who vanquishes all others is termed the alpha male. For the alpha male, »might is right.«

Humans, by contrast, do not have alpha males. In small-scale societies of humans there tend to be no leaders, such as among nomadic hunter-gatherers and forest horticulturalists. In larger groups human societies can have leaders, but unlike the arrangement among non-human primates, human so-called leaders do not personally defeat all their subordinates in one-on-one fights. Human leaders achieve their high ranking positions through the support of an alliance. If the leader's allies cease to support him, he will lose his top position.

The fact that hunter-gatherers mostly have no leaders is particularly important given that their societies provide strong indications of the nature of human society in the Late Pleistocene (Boehm, 2012). The lack of alpha males in nomadic hunter-gatherers therefore indicates that by the late Pleistocene, natural selection in humans no longer favored the alpha-male style of behavior. The question is why. What had happened among the males of human ancestors to suppress or modify the »ubiquitous drive toward individual selfishness«?

2. Reduction of competition among males in humans

In an extensive review of small-scale societies Boehm (1999) found occasional cases of individual men using their fighting ability to attempt to dominate other men by threats, bullying or murder. When this happened, Boehm reported, the community would first