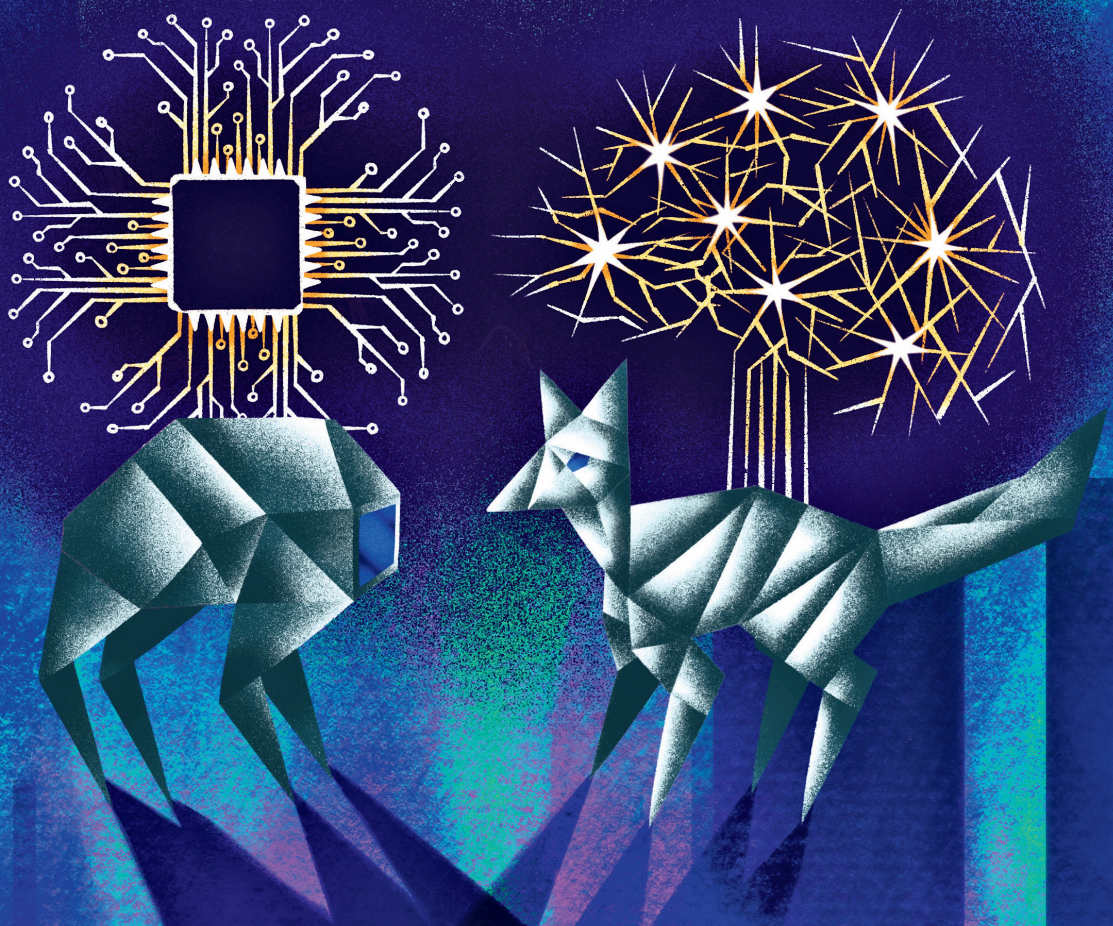
ACTA UNIVERSITATIS GOTHOBURGENSIS
ACTA PHILOSOPHICA GOTHOBURGENSIA 41

Nonhuman Moral Agency

A Practice-Focused Exploration of Moral Agency in Nonhuman Animals and Artificial Intelligence

Dorna Behdadi



UNIVERSITY OF GOTHENBURG

Nonhuman Moral Agency

ACTA UNIVERSITATIS GOTHOBURGENSIS
ACTA PHILOSOPHICA GOTHOBURGENSIA 41

Nonhuman Moral Agency

A Practice-Focused Exploration of Moral Agency in Nonhuman Animals and Artificial Intelligence

Dorna Behdadi

UNIVERSITY OF GOTHENBURG

ISBN 978-91-7963-149-9 (print)
ISBN 978-91-7963-150-5 (pdf)
ISSN 0283-2380

The publication is also available in full text at:
http://hdl.handle.net/2077/78610

Subscriptions to the series and orders for individual copies sent to:
Acta Universitatis Gothoburgensis
PO Box 222
SE-405 30 Göteborg, Sweden
or to acta@ub.gu.se

This research was supported by the Swedish Research Council (VR), contract no. 2014-40, for the Lund Gothenburg Responsibility Project.

Cover Art: Eli Leo Ydén

Photographer: Monica Havström

Print: Stema Specialtryck, Borås, 2023

Abstract

Title:	Nonhuman Moral Agency: A Practice-Focused Exploration of Moral Agency in Nonhuman Animals and Artificial Intelligence
Author:	Dorna Behdadi
Language:	English
ISBN:	978-91-7963-149-9 (print)
ISBN:	978-91-7963-150-5 (pdf)
ISSN:	0283-2380
Keywords:	moral agency, moral responsibility, artificial intelligence, nonhuman animal, practice-focused, moral status, moral patient, blame, moral psychology, Strawson, consciousness, participant stance, social norm

Can nonhuman animals and artificial intelligence (AI) entities be attributed moral agency? The general assumption in the philosophical literature is that moral agency applies exclusively to humans since they alone possess free will or capacities required for deliberate reflection. Consequently, only humans have been taken to be eligible for ascriptions of moral responsibility in terms of, for instance, blame or praise, moral criticism, or attributions of vice and virtue. Animals and machines may cause harm, but they cannot be appropriately ascribed moral responsibility for their behavior.

This thesis challenges the conventional paradigm by proposing an alternative approach where moral agency is conceived as the competence to participate in moral responsibility practices. By shifting focus from intra-individual to contextual and socially situated features, this *practice-focused* approach appears to make the attribution of moral agency to nonhuman animals and AI entities more plausible than commonly assumed.

Moreover, considering the current and potential future prevalence of nonhuman animals and AI entities in everyday settings and social contexts, a potential extension of moral agency to such entities could very well transform our social, moral, and legal practices. Hence, this thesis proposes that the attribution or withholding of moral agency to different entities should be carefully evaluated, considering the potential normative implications

List of Papers

I. Behdadi, D., & Munthe, C. (2020). A Normative Approach to Artificial Moral Agency. *Minds and Machines*, *30*(2), 195-218. doi.org/10.1007/s11023-020-09525-8

II. Behdadi, D. (2021). A Practice-Focused Case for Animal Moral Agency. *Journal of Applied Philosophy*, *38*(2), 226-243. doi.org/10.1111/japp.12486

III. Behdadi, D. (submitted for publication). The Moral Claimant Account of Moral Agency.

IV. Behdadi, D. (submitted for publication). Moral Patiency Grounds Partial Moral Agency.

To Kazim and Baba

You are with me always

Acknowledgments

I would not be standing at the finish line (although I have yet to cross it when writing this) were it not for all the amazing people who have helped and supported me during these past years.

First of all, Elin *Belin* Pernevik, my love, I am eternally grateful for your relentless support. Your love, encouragement, and patience have kept me going through it all. Not to mention, you have more than anyone had to suffer what it means to live with a philosophy PhD student: Enduring early-morning thought experiments, late-night discussions on the mind-body problem, and even amateur moral psychological experiments, you have weathered it all with grace. I am incredibly fortunate to have you by my side, and you truly mean the world to me.

To my dear mother, Shahnaz Haghshenas, *Mami*, thank you for all the thought-provoking discussions and your openness to new perspectives and ideas. I clearly inherited my curiosity and my inquisitive traits from you. From where I stand, you are the original philosopher of our family.

Sarvi Glemfors and Delli Holmblad, my dear big sisters, thank you for being there through it all. I have always wanted to make you proud.

Abbe Balkhi and Taghi Moradi, I am lucky to have gained such kind and smart bonus brothers. I am glad you make up for my lack of technological know-how.

When I defend my thesis a couple of months from now, there will inevitably be two family members missing from the audience. During my doctoral studies, I lost, first, my twin, Kazim Behdadi, and later my father, Parviz Behdadi.

Kazim, you are forever my clone and best friend. The hole you left in the world is astronomical. I will never be the same without you. But I will try to live up to your ideals.

Baba, my greatest supporter. I miss your gentle love and care. I am certain that my fondness for all things tender and poetic comes from you.

Bereavement, especially when sudden, forces you into an almost existential crisis. Up is down, light is dark, and you are suddenly and unwillingly flung into a reality that is forever broken. I was not sure how, or if, I would make it through. But I did. And there are many people in addition to my family responsible for helping me do so. Eli Leo Ydén, Halima Handulleh, Tony Jageteg, Ebba Ewaldh, Sofia Skoglund, Lovis Thedéen, Jonna Håkansson, Ditte Enhorn, Anahita

Amirpour, Mikael Karlsson, Marty Francis Repka, thank you for being there when everything turned dark.

Of course, this thesis would not have been finished without the support and feedback from my colleagues at the Department of Philosophy, Linguistics and Theory of Science at the University of Gothenburg.

Christian Munthe and Erik Malmqvist, dear supervisors, I would not be writing these words in the first place were it not for you. Christian, I am very lucky to have had you as my supervisor. Your never-ending enthusiasm rubs off, and you have an almost uncanny ability to make people see things constructively, no matter the issue or topic. Moreover, you could always tell when to push me and when to tell me to go easy on myself. Erik, I am very happy to have had you as my supervisor. Although it has sometimes been difficult for you to get a word in and for me to hear you during supervision meetings (yes, I'm looking at you Christian!), your knowledge and counsel have improved this thesis countless times over. Thank you also for always being so kind and encouraging. I am deeply grateful to you both for bearing with my elliptical writing, weirdly placed commas, and many anxious questions during these years. I will truly miss having you by my side.

Bengt Brülde, thank you for your valuable guidance and feedback on paper drafts and other things PhD during my initial years as a doctoral student. And thank you for pointing me toward Zen meditation when I was at my lowest. Sitting helped me get through the initial stages of grief.

Joakim Sandberg, Bengt, and Christian, you might not know it, but your encouragement early on at the bachelor's level planted the seed of pursuing a doctoral degree in philosophy. Thank you for helping me believe in myself!

Sofia Jeppsson, Per-Erik Milam, and Benjamin Matheson, thank you for being such smart and approachable colleagues during my first two years and for your feedback on my paper drafts.

Petra Andersson, thank you for showing me that philosophers can be rebels too. I had a lot of fun setting up and teaching the animal ethics course with you!

Leila El-Alti, thanks for being such a fun and smart office buddy at Olof Wijksgatan and thank you for putting together the dissertation checklist. It really came in handy!

M. Hadi Fazeli, Alexander Andersson, John Eriksson, Ragnar Francén, Thomas Hartvigsson, Jasmine Elliott, Richard Endörfer, Olle Blomberg, Davide Fumagalli, Yuliya Kanygina, Paiman Karimi, Ida Hallgren, Georg Schmerzeck, Fausto Corvino, Gunnar Björnsson, Paul Russel, Mattias Gunnemyr, Kristin

Mickelson, Robert Hartman, Marco Tiozzo, Nina van Heeswijk, Louis Larue, Clément Fontan, Niels Nijsingh, Stellan Welin, Anna-Sofia Maurin, Anders Tolland, Ylwa Sjölin Wirling, Frans Svensson, Susanna Radovic, and others (whom I may have inadvertently omitted), thank you for your valuable input on my paper drafts and for all the interesting discussions during philosophy research seminars.

Needless to say, I am also grateful to many non-philosophers at the department.

Thank you Monica Havström for always making me feel welcome and for making me look professional in photos.

Thank you, Lena Eriksson, Emelie Seeliger, Johan Söderberg and Christopher Kullenberg for your valuable advice, guidance and kind support when I was not feeling well.

Thanks also to current and past key personnel at the department for your assistance during these past years: Fredrik Engström, Alexander Almér, Ulrik Sixt, Jennifer Stråle, Hanna Edblom, Cecilia Groglopo, Andreas Ott, Peter Johnsen, Robert Adesam, Martin Jacobsson, Mikaela Agri, and Helena Bjärnlind.

A special acknowledgment as well to the brilliant peeps who are part of the *AI in Society* course planning and reading groups at the department: Hadi (the initiator!), Bill, Eleni, Simon, Ellen, Chris, Asad, Bahareh, Nikolai, and others.

And a shoutout to all the other bright and delightful individuals who make lunch breaks at the department enjoyable: Aram, Dominik, Orvar, Vidya, Alex B., Tjeerd, Giacomo, Amandine, Hektor, Doris, Niclas, Dick, and everyone else!

Of course, I have also received support from people outside of the department.

Pär Sundström, thank you for teaching me much of what I know about the philosophy of consciousness and for your encouragement in the early days before I applied for a PhD student position.

Matthew Talbert, thank you for your valuable feedback during and after my final seminar. You truly helped improve this thesis.

A big thanks to Karl Pettersson at ACTA for helping me with everything related to getting my thesis ready for the printer!

Olle Häggström, I have lost count of how many AI-themed panels we have participated in together. Thank you for the fun talks and your relentless effort to raise awareness about the risks of artificial general intelligence.

Thank you also to Jonna Håkansson, Juan Velasquez, Thomas Laurien, Helena Pedersen, and other former and current members of Gothenburg University's network for Critical Animal Studies in the Anthropocene for all the great seminars

and talks organized together over the past five years. I am happy to have found such great scholars dedicated to raising awareness about our exploitation of nonhuman animals.

Thank you as well to the organizers of the Animal Morality Conference in Vienna and the Inaugural Meeting of the Philosophy of Animal Minds and Behavior Association for inviting me to speak and making my last spring as a PhD student the very best.

Thank you also to the organizers of TEDxGöteborg 2019 – *Disrupting Status Quo* for inviting me to speak about AI at your event. I learned loads about presenting at your coaching sessions that I will have use for many years to come!

I would likewise like to express my gratitude to Filosofiska fakulteternas gemensamma donationsnämnd and Kungliga och Hvitfeldtska stiftelsen. Your generous funding helped me extend my studies and finish my thesis.

Eli Leo Ydén, thank you for the stunning cover illustration for this thesis and for bearing with my many requests during the drawing process!

When finishing this section, I am in Manchester competing in the roller derby tournament *Eurocup*. I therefore want to take the opportunity to thank my current and former teammates in Gothenburg Roller Derby, Malmö Crime City Rollers, and Dock City Rollers for keeping me grounded and sane these past years. Sitting in front of a screen or a book by yourself for hours on end can take a toll on your mind and body. Getting my regular share of adrenaline and hits on the track has provided a much-needed break. I am immensely grateful for being part of the derby community!

I also want to thank the employees at BeActive Urmston sports venue, Prestwich, Manchester, for letting me use your staff room to finish this thesis during the last day of Eurocup 2023.

Last but certainly not least, I want to thank some of the nonhuman animal persons in my life. I will begin with those who are no longer here. Max, thank you for teaching a young Dorna in your own *finchy* way to have compassion for and defend sentient beings no matter their species. Pedro, my feline brother from another mother, thank you for letting me share my life with you. I still wish we would have had more years together. Mir, Boye, Hafez, and Shirin, thank you for the fun and cuddles. And for letting me get a glimpse into the inner workings of degus (some of which is printed in my bachelor's thesis in zoology). I hope you are all munching

on all the seeds and rolled oats you can wish for. Pelle, dear cat-gramps, thank you for being the calm presence I needed when times were tough. I miss you.

Finally, I would like to express my gratitude to my current nonhuman animal family members: Manoosh, my dear shy brave friend. Thank you for teaching me that trust is earned and never forced or hurried. Benni, sweet goofball, your biscuit making love is priceless. Dear Minou, my sensitive, smart and compassionate village pup. I do not think I would be on my current path were it not for the lessons on tact, patience and consent you have taught me.

Dorna Behdadi, Manchester, 2023

Contents

1 Introduction

The general topic of this thesis is the moral agency of nonhuman beings, such as nonhuman animals and machines and software that use artificial intelligence. The overarching questions asked are: Can moral agency be ascribed to nonhuman entities? If so, in what sense, or to what extent, can moral agency be ascribed to them? *Should* it be so ascribed? What criteria and boundaries are valid for affirming or denying the moral agency of nonhuman beings? I engage with these questions via four research papers (I-IV) and this introduction of six chapters.

Traditionally, philosophers have approached such questions (also when asked about humans) by applying intuitive assumptions of moral agency as requiring certain given intraindividual capacities related to, for example, freedom or deliberate reflection. However, this thesis develops an alternative approach that primarily understands moral agency as a socially situated and contextually contingent competence or skill. This alternative conception is developed and applied to possible cases of nonstandard moral agents, where nonhuman moral agents, like animals and artificial intelligence entities, constitute the central but not only examples.

My treatment of these questions links to general longstanding themes in contemporary philosophical discourse about moral agency and responsibility, debates regarding the fringes and boundaries of human moral agency, as well as specific more recent discussions concerning the possibility of nonhuman moral agency.

Recent years have seen an increasing interest in issues relating to the boundaries of moral agency. Philosophers and others are asking questions about the possibility of moral agency outside the scope of typical adult humans, the often-assumed paradigm of a moral agent. Many of these authors are discussing where to draw the line *between* members of our own species. Can, for instance, young children, or people with alleged "disorders of agency" (Pickard & Ward, 2013, p. 1134) be morally, as opposed to merely causally, responsible for their actions (Shoemaker, 2015; Pickard, 2011, 2014, 2015, 2017; Pickard & Ward, 2013, Kennett, 2002, 2009; El-Alti, 2023, Burroughs, 2020)?

Others, myself included, are (also) interested in the possibility of moral agency beyond the classificatory space of Homo sapiens. Can nonhuman animals, like dogs or apes, act morally rightly or wrongly, well, or badly (de Waal, 2006; Bekoff & Pierce, 2009; Rowlands, 2012; Vincent et al., 2018)? If so, can they be praise- or blameworthy for their conduct? And can artificial intelligence entities, like autonomous robots or advanced computers, behave in morally right or wrong ways, and ever be properly considered praise- or blameworthy for their behavior (Himma, 2009; Johnson, 2006; Floridi & Sanders, 2004; Sullins, 2011; Parthemore & Whitby, 2013)?

1.1 Background

It is, perhaps, not difficult to see why questions about the boundaries of moral agency are receiving increasingly scholarly attention. For one, the general philosophical discussion about moral agency and responsibility has recently become less and less concerned with questions about free will and conscious deliberation, and more so with the nature of those of our everyday interactions and practices that assume moral agency and responsibility. Hence, the (quite idealized) picture of moral agents as free, independent, and rational has found competition in ideas that emphasize the social and emotional significance of our practices of attributing or withholding moral responsibility. Attempts at describing the nature, shape, and function of said emotions and practices have, in turn, made apparent the need for nuanced, flexible, and sensitive conceptions of moral agency.

At the same time, empirical psychological data indicate that the cognitive processes underlying our behaviors, evaluations, and decisions are often not transparent to us, and diverge from the reasons we ourselves provide (in the form of, for instance, explanations, excuses, or justifications). For example, when people are asked to provide reasons for their choices and evaluations, they regularly seem to lack awareness of the stimuli that influenced their responses, and instead point to other (ad hoc) reasons (Nisbett & Wilson 1977; Haidt, 2001; Wilson, 2002; Doris, 2002; Bargh & Ferguson, 2000; Bargh & Chartrand 1999; Bargh et al, 2012; Kahneman, 2011).

Psychological states and processes are thus to great extent non-conscious and determined by circumstantial and social factors, all of which implies that even moral attention and motivation, and subsequently, our moral judgments and decisions, are influenced by factors that are typically unknown to us and often independent of conscious guidance (Bargh & Ferguson, 2000; Wilson, 2002;

Kahneman, 2011; Doris, 2002). For instance, experiments have shown that people's tendency to help strangers in distress is largely dependent on (morally irrelevant) situational factors, such as having found or not found a coin (Isen & Levin, 1972; Darley & Batson, 1973).

Furthermore, research has also revealed the extent to which evaluations, behavior, and decisions are influenced by implicit biases (Uhlmann & Cohen 2005; Greenwald & Krieger, 2006; Nosek et al. 2007). Taken together, these results challenge the relevance and validity of traditional philosophical conceptions of moral agency which assume free will or control, robust character traits or values, and rational or conscious deliberation. They instead, lend support to accounts that favor moral agency as, at least in part, dependent on, and driven by, habitual and non-conscious processes as well as circumstantial and social factors. Many recent accounts, therefore, favor conceptions of moral agency that are better suited to account for the nuances and intricacies of real-life interactions where moral agency is assumed (Strawson, 1962/1982; Wolf, 1987/2013; McKenna, 2012; Wallace, 1994; Watson, 1987/2004; Shoemaker, 2015; Sie, 2014; Arpaly, 2002; Doris, 2015; Vargas, 2013; McGeer, 2019; Macnamara, 2015a).

Above and beyond these developments, the recent expansion of the interest in nonhuman moral agency specifically seems to have been inspired by additional factors. For instance, new findings in comparative cognition have challenged traditional assumptions about the behavior and minds of nonhuman animals (Allen & Bekoff, 1997; Shettleworth, 2012; Beran et al., 2014; Andrews, 2020a; Andrews & Monsó, 2021) as wholly distinct from, and lesser-than, human. This "cognitive turn" (Jamieson & Bekoff, 1992, p. 110) in animal science has, at the same time, been accompanied by the emergence of multidisciplinary research fields within the humanities, social and educational sciences, such as human-animal studies (or anthrozoology) (Shapiro & DeMello, 2010; DeMello, 2021), and critical animal studies (Best et al., 2007; Best, 2009).

This *animal turn*[1] in the humanities and social sciences marks a significant change of focus (and, according to some, even a change in method and theory (Pedersen, 2014) in the academic study of other animals. Instead of merely seeing and describing nonhuman animals as entirely alien or other *objects* to study or observe, the starting point is that they are possible *subjects*.

In parallel, discussions on the moral and political standing of animals have increasingly started to apply concepts such as justice, personhood, and citizenship

[1] Simmons and Armstrong (2007) accredit the phrase *animal turn* to Sarah Franklin who used it at the 2003 Cultural Studies Association of Australasia conference. Also see Pedersen (2014).

(Nussbaum, 2005; Varner, 2012; Kymlicka & Donaldson, 2014; Sunstein, 2005; Aaltola, 2008; see also Cochrane et al., 2018), in contrast to the heavy emphasis of earlier discussions on sentience, protection and basic interests or rights (Singer, 1975; Regan, 1983/2004). In addition, the last decade has seen an increased scholarly interest in and recognition of nonhuman animals as co-inhabitants, participants, partners, co-workers, or even citizens (Donaldson & Kymlicka, 2011; Blattner et al., 2019; Meijer, 2019) of human societies and political and social institutions. Together, these developments reflect a trend where scholars are moving away from viewing nonhuman animals as mere passive recipients or objects of care and concern towards viewing them as potential social, political, and even moral *agents*.

At the same time, the development and potential of increasingly advanced and autonomous machines and programs has raised a number of philosophical, ethical, and legal issues. Before the twenty-first century, most philosophical interest in the possibility of artificial intelligence, consciousness, and morality was raised in science fiction literature and drama. Prominent examples can be found in Isaac Asimov's short story "Runaround" which introduced the "Three Laws of Robotics" (1942), and Philip K. Dick's "Do Androids Dream of Electric Sheep?" (1968), which, among other things, raised epistemic and ethical issues regarding the possibility of consciousness in artificial humans (or *androids*). The comparative lack of academic philosophical interest can probably be attributed to the technological limitations of earlier AI systems (see Franklin, 2014).

However, the last two decades have seen a dramatic increase in the development and use of AI. AI systems are now being deployed in a wide range of contexts, such as education, healthcare, transportation, surveillance, finance, and the military (see Coeckelbergh, 2020a; Stone et al., 2016). Some examples of recent AI-based applications that have spurred public, as well as philosophical, interest, and controversy are Open AI's chatbot ChatGPT and their image generator DALL-E (Floridi, 2023; Rudolph et al., 2023).

These developments have been accompanied by the emergence of a broad field of research into the philosophy and ethics of artificial intelligence and robotics.[2] The questions of this field can be divided into two main types: one regarding the behavior and responsibility of *humans* in designing, using, and treating AI systems[3] and the other concerning the behavior, nature, and (moral) responsibility of

[2] This is called *the ethics of artificial intelligence* (Müller, 2021).

[3] This is where *robot ethics* or *roboethics*, which concerns the design, use, and treatment of robots, belongs (Veruggio, 2006; Müller, 2021).

machines that use AI (Bostrom & Yudkowsky, 2018; Coeckelbergh, 2020a).[4] Much of the focus within this latter branch is still concerned with questions about how to design and implement AI systems to ensure that they behave in accordance with certain human moral standards. Is it, for instance, possible to build and program autonomous vehicles, such as self-driving cars, that can not only move safely in traffic but also make decisions based on pre-programmed or learned moral rules or principles? A related, but further, question, and one that I address in this thesis, is whether machines (including robots) that use AI can themselves be considered moral agents.

Artificial intelligence entities and nonhuman animals both represent categories of nonhuman others. That is, they are types of beings or entities who fall outside the notion of the typical adult human, the presumed paradigm of a moral agent. As such, AI and animals are nonhuman nonstandard cases of possible moral agents. Yet, they each bring their own set of questions to a head. Advanced and autonomous machines and programs are artificial and man-made, and many times capable of written or spoken communication. However, while AI may be described as, and considered, intelligent, as well as possible to be engaged with linguistically, there is widespread skepticism about the existence and possibility of *conscious* machines (Purves et al., 2015; Dehaene et al., 2017; Johnson & Verdicchio, 2018; Birhane & van Dijk, 2020).

Nonhuman animals, on the other hand, are natural organisms and generally considered to possess phenomenal mental states[5] such as conscious sensations and feelings.[6] However, generally speaking, other animals are usually not attributed high (human) levels of intelligence (Nakajima et al., 2002), such as rationality or (self)reflection and (certain forms of) self-awareness (Korsgaard, 1996; Tulving, 2005) and are presumed incapable of language, and sometimes also beliefs (Davidson, 1982, 1984) and concepts (Davidson, 1982, 1984; Laurence & Margolis, 2012). In this way, AI and nonhuman animals present differential sets of similarities and differences in relation to the (perceived) features of typical adult humans.

The assumptions and perceived characteristics associated with each of these groups therefore prompt different intuitions and produce slightly different areas

---

4 This is sometimes referred to as *machine ethics* (Anderson & Anderson, 2011; Moor, 2006; Allen et al., 2006).

5 At least in the case of vertebrates and certain invertebrates (Low et al., 2012).

6 A contemporary exception is found in Carruthers (2019), who denies that other animals possess phenomenal consciousness.

of inquiry about the nature, requirements, and possibility of moral agency. In this way, considerations of moral agency in animals and machines can also serve to make explicit assumptions about human nature and human uniqueness, as well as about (moral) agency and responsibility in general (Loughnan & Haslam, 2007).

At the same time, there is some reason to question these assumed differences between animals and machines, both concerning their presumed grounds and their related implications. For one, the artificial-natural distinction is, in and by itself, far from clear. Moreover, it is not obvious what relevance any such distinction would have for questions relating to moral agency, such as questions concerning phenomenal consciousness, intentionality, conscious deliberation, and free will. While machines are man-made, one can imagine advanced AI capable of adjusting to environmental changes in flexible ways similar to the biological and behavioral adaptation and learning of natural organisms. In addition, if machines were manufactured from biological materials like those of sentient organisms, there appears to be room to question the assumption that they cannot be sentient due to their mere physical constitution. Similarly, artificial selection represents an ancient albeit *directed* process in which humans exploit (animal) breeding to develop living and sentient beings with certain characteristics. Likewise, more recent biotechnology, such as artificial insemination, IVF, and cloning, have made it possible to create both human and nonhuman animals without relying on certain aspects of *natural reproduction*. Future biotechnological developments may make it possible to create organisms in ways that even more radically circumvent naturally available reproductive means. The assumed distinction between the natural and the artificial may be further blurred in light of increasing integration of natural organisms and non-organic technology, as in the case of prostheses, implants, or artificial organs, aimed at restoring or enhancing physical or mental capabilities.

The question of to what extent we have reason to cling to distinctions between the natural and the artificial, and related more advanced ones will be discussed further on in this thesis. For now, however, I will proceed from the assumptions in these debates regarding the difference between artificial and organic agents, regardless of who exactly fits either category.

One issue commonly discussed in relation to the possibility of artificial moral agents is whether an artificial entity can have the type of freedom or control traditionally assumed to be required for moral agency, seeing that it has been designed and programmed to behave in certain ways. If the answer to this question is no, the obvious follow-up seems to be to ask what this might mean for the possibility of human moral agency. Is there, for instance, any relevant difference

between the ways machines are determined by design and programming, and the various genetic and environmental conditions that determine the choices and behaviors of human beings?

Another common worry is that, despite the seeming linguistic evidence of mental states such as beliefs and attitudes in some artificial entities, these displays cannot be considered relevant for moral agency, as they are mere simulations and not the real deal. However, as discussed in Paper I, this type of epistemic skepticism seems to raise similar issues in the case of humans.

Considerations about nonhuman animal moral agency make explicit other fundamental intuitions and assumptions. Is it, for instance, sufficient that an animal acts or behaves in accordance with what is right or good for them to be considered a moral agent? Or does moral agency require more than a *benevolent disposition*, such as, say, conscious and rational deliberation? If so, what are the implications of such requirements for typical adult humans? To what extent is, for example, the behavior and choices of humans typically taken to indicate moral agency and responsibility really grounded in conscious reflection, rather than habitual, affective, or non-conscious processes?

Artificial intelligence entities and nonhuman animals thus present us with distinct sets of similarities and differences to (typical adult) humans. As such, considering these cases has the potential to make apparent different assumptions and issues regarding moral agency in general. Considering the possibility of nonhuman moral agency is therefore akin to, and can complement, discussions about the possibility of moral agency in nonstandard or *marginal* human cases, such as very young children or adults with allegedly moral agency-*undermining* conditions or disabilities (Shoemaker, 2015; Pickard, 2011, 2014, 2015, 2017; Pickard & Ward, 2013, Kennett, 2002, 2009; El-Alti, 2023, Burroughs, 2020)?

Considerations of moral agency in nonhuman entities are thus philosophically significant because they require us to spell out, and critically reflect on, the grounds of moral agency in general, thereby expanding and enrichening the general philosophy of moral agency and responsibility. In other words, we are forced to face the question: What does it take for any entity to be a moral agent? Critical assessment may force us to reconsider the validity and relevance of commonly assumed requirements of moral agency. We may, for instance, find that some requirements exclude some people whom we, in fact, usually include in our moral responsibility practices. Or, worse, we may find that some of our commonly assumed requirements are, on reflection, too strict to include *any* human. But the inquiry into nonhuman moral agency is also of importance for more practical reasons. The

practices and behaviors of an ever-increasing human population have had (and still has) a massive impact on the living conditions of nonhuman animals. We share our homes and lives with dogs, cats, horses, and other domestic animals, and have done so, in some way or another, for thousands of years (Jensen, 2017). Some consider animals used for work as partners or colleagues (Dashper, 2016; Charles et al., 2022), and many people view their companion animals as friends and even family members (Fox, 2006; Charles & Davies, 2008).

In addition, human settlements, like cities and other urban areas are co-inhabited by so-called *liminal* animals, like brown rats, foxes, squirrels, rabbits, raccoons, rabbits, hares, jackdaws, pigeons, and magpies. These animals are neither fully wild nor fully domesticated[7] but exist in a marginal space, rely on anthropogenic food sources, such as human waste, take shelter in human-made structures, and are more or less used to the presence of humans (Donaldson & Kymlicka, 2011, 2016; Kalof & Whitley, 2021; Brouwer, 2018).

Nonhuman animals, such as mice, rats, zebrafish, rabbits, guinea pigs, chickens, dogs, cats, reptiles, amphibians, monkeys, and other primates, among others, are also kept and used by humans for scientific purposes (Hickman et al., 2017; ALURES, 2020). The number of animals used in experiments worldwide has been estimated to be close to a staggering 200 million individuals just for the year 2015 (Taylor & Alvarez, 2019).

Not to forget, hunting and farming have caused wild mammal biomass to decrease by 85% through direct killing and habitat loss. At the same time, livestock, that is, animals bred, confined, and slaughtered for human consumption, have come to outweigh wild mammals and birds by a factor of ten. Mammals used as livestock, such as cattle, pigs, sheep, and goats, now make up 94% of global mammal biomass (excluding humans), and poultry make up 71% of the world's bird biomass. In fact, humans and our livestock together outweigh the biomass of all land-dwelling vertebrates combined (Ritchie et al., 2022; Ritchie, 2019; Bar-On et al., 2018).[8]

At the same time, there is an accelerating development of increasingly intelligent, autonomous, self-learning, and socially capable and integrated machines and programs. People interact with, and are increasingly exposed to, AI on a regular basis through, for instance, applications using speech recognition (like personal assistants and customer service) (Xu et al., 2020), social media bots

---

7 Although domestic animals can be liminal, such as feral or stray cats, dogs, and rabbits.

8 These human-induced effects are also part of and cause of what is sometimes referred to as the *Anthropocene*, the "human-dominated geological epoch" (Lewis & Maslin, 2015, p. 171).

posting, for example, political content (Stieglitz et al., 2017), non-playable characters (NPCs) in video games (Mehta et al., 2022n), recommendation and targeting systems on streaming services, online stores or social media (Milano et al., 2020), digital or robotic teachers or teaching assistants (Kim et al., 2020), companions and robotic pets (Skjuve et al., 2021; Petersen et al., 2017; Aarskog et al., 2019), and medical assistants and caregivers (Morley et al., 2020).

What is more, people seem to readily anthropomorphize artificial entities that appear and behave in humanlike and empathetic ways (Pelau et al., 2021; Waytz et al., 2014). We treat robots in ways indicating the implicit attribution of mental states to them, even when verbally denying that they have such states (Thellman et al., 2020). And some people even view and treat robots and AI systems as coworkers (Sauppé & Mutlu, 2015) or appreciated colleagues (c.f. Nyholm & Smids, 2020), friends (Newman, 2014; see also Elder, 2017; Skjuve et al., 2021), romantic companions or sex partners (Döring et al., 2020). While some such attribution may be unintentional from the perspective of designers and programmers, many AI applications are designed to exploit human cognitive and affective biases and target vulnerabilities (Nadler & McGuigan, 2018)[9]

Considering the current, as well as potential future, prevalence of nonhuman animals and AI in human social contexts and settings, a potential expansion of who are and who are not included as presumed moral agents could very well transform the basic building blocks of our social, moral, and legal practices. Moral agents are usually considered eligible for certain kinds of responses and treatments, such as blame and praise, ascriptions of moral responsibility, and other social sanctions, all of which would otherwise be inappropriate. The inclusion or exclusion of artificial entities and nonhuman animals as moral agents may therefore present further and potentially far-reaching practical implications. In addition, in light of a possible link between moral agency and moral patiency, these questions may also reveal further and quite pressing normative issues. For instance, I suggest that, due to our moral psychological inclinations, we have normative reasons for ascribing to moral patients a particular type of moral agency (Chapter 5).

[9] This is not because designers and programmers necessarily wish to mislead or deceive users. Instead, the reliance on cues that elicit *humanness* mediate interaction and trust, and subsequently serve to improve function and efficiency of many AI applications (see, for example, Hancock et al., 2011).

1.2 The Rest of this Thesis

Each of the papers (I-IV) deals with one or several of the themes mentioned in the preceding section. Paper I focuses on charting and analyzing the various questions discussed in relation to artificial moral agency, such as whether free will and consciousness are necessary requirements. The conclusion of this paper is that much of the debate has been stuck in theoretical disagreements of little to no practical significance, which highlights the need for an alternative approach. Paper II picks up on this suggestion, by investigating the possibility of nonhuman animal moral agency using such an alternative approach. If moral agency is defined in terms of participation in certain social practices, I argue, the prospect of nonhuman animal moral agency appears to be more likely than usually thought.

Paper III employs a communicative understanding of moral responsibility practices, specifying participation in such practices in terms of engagement in a certain type of moral exchange or conversation. Given a functional analysis of these communicative practices, moral agency is then claimed to be applicable to typically exempted humans and nonhuman entities in virtue of them qualifying as a particular kind of participant. Paper IV draws on finding from Paper III and argues that there are independent normative reasons for avoiding the exemption of moral patients from moral agency. We should, as a default, try to view all moral patients as (potential) sources or makers of moral claims and demands. Each of these papers are summarized and elaborated on further below.

In addition to this preface, and the four papers just summarized, the thesis consists of four chapters, devoted to describing and analyzing central topical themes of the thesis. The over-arching aim of these chapters is to provide a philosophical background to the papers and to better explain their philosophical contributions. Chapter 2 clarifies some central philosophical concepts used in discussions about moral agency and responsibility and outlines some influential contemporary accounts or positions about moral agency and responsibility of importance to the claims and arguments developed in this thesis. Chapter 3 continues by specifying the underlying strategy and possible virtues of the favored method and account of this thesis: the practice-focused approach to moral agency. The chapter proceeds by discussing some prominent examples that fit the description of this approach, as well as contributions made to this approach in this thesis.

Chapter 4 discusses and analyzes the various debates on nonstandard human and nonhuman cases of moral agency, points out potential gaps and limitations in

these discussions, and highlights possible remedies suggested in this thesis. Chapter 5 addresses the relationship between moral agency and moral patiency by assessing their traditionally assumed differences and proposing a revision in light of the normative implications of Paper IV. The thesis is then concluded by an overview of its main findings and conclusions (Chapter 6) and extended summaries of the Papers I-IV (Chapter 7).

2 Moral Agency and Moral Responsibility

A standard way of defining moral agency is to say that someone is a moral agent to the extent that they can be ascribed the moral qualities of their actions, so that they are not only performing, say, a wrongful act but thereby also, in virtue of this, a wrongdoer. This links closely to a notion of moral responsibility according to which all moral agents are, in virtue of this agency, potentially blame- or praise-worthy for their acts (in view of the acts' moral qualities) or they are fit to be ascribed forward-looking responsibilities, in terms of duties (Talbert, 2022).

However, none of this really answers what it takes to be a moral agent or, in effect, to be morally responsible. Neither does it explain how these concepts are linked to each other, given specific ascriptions of agency and responsibility, or what more exact actions and moral reactions might be appropriate or justified given that someone is or is not a moral agent or morally responsible. Thus, to approach the question about moral agency in nonhuman entities and nonstandard human cases, we need more detailed and specific theorizing.

I believe that a fruitful way of approaching this challenge is to attend to the social contexts and practices to which these concepts apply, and in which they are used, for instance by ascribing aspects of moral agency to some entities while withholding them from others. This practice-focused approach, as I will call it, to the question of moral agency can be summed up thus: to be a moral agent is to participate in moral responsibility practices. This approach forms a fundamental part of the methodology of this thesis and is developed and defended in Chapter 3. What is included in moral responsibility practices is discussed at length in Papers II and III.

For now, however, I will use the loose characterization made above as starting point for introducing the challenges involved in making sense of the concepts and issues mentioned at the outset of this section. From my standpoint of a commitment to the practice-focused approach, a good place to start elucidating these concepts and issues is to consider some everyday scenarios involving examples of

paradigm moral agents, as well as beings typically exempted from moral agency in our moral practices.

Let us start with two, seemingly similar, scenarios, which both involve two human beings but seem to evoke very different intuitions (and correlated action tendencies) regarding moral agency. In the first scenario, a baby reaches toward someone's face and pokes them in the eye, making it sore and red. In the second scenario, a (typical) adult human reaches toward a person's face and pokes them in the eye, causing the same outcome. When considering the first scenario, we may indeed say that the baby *caused* the eye to become sore and red. In other words, the baby can be properly considered *causally* responsible for the afflicted person's pain and discomfort. This is because the baby is clearly the salient cause of the eye getting sore. We might even say that the baby poked the person's eye *intentionally*. They might, for example, have formed a (rather skillful, however unfortunate) habit of aiming for, and successfully poking, people in their eyes.

Still, it would seem misplaced, and even cruel, to judge the baby as responsible in a moral sense – by implying that their behavior discloses a bad character, by trying to morally reason with them, or blaming them. However annoying, troublesome, or inconvenient the actions of a baby may be, babies do not seem to be eligible for ascriptions or reactions of moral responsibility. Following Shoemaker's (2011, 2015) tripartite theory of moral responsibility, one could thus say that babies are ineligible for any of the three types of responsibility assessments in terms of attributability, answerability and accountability.[10] Babies are not eligible for moral assessments about how the moral features of the behavior, character trait, or its consequences could be attributed in the right way to the baby's value system or moral character (Scanlon 1998; A.M. Smith 2005; Hieronymi, 2008). Neither do babies appear to be eligible for answerability demands, which would involve, that is, asking or urging the baby to provide reasons for their behavior and assessing the baby's evaluative judgments (Shoemaker, 2011, 2015). Nor do babies appear to be open to being *held* responsible for their actions, for example, by demanding certain conduct from them and blaming them if they fail to comply (Watson 1996/2004; Shoemaker, 2015). Neither causal nor intentional agency thus seem sufficient for *moral* agency and responsibility.

While babies do not seem to inhabit the position, role, or status required for the appropriate application of concepts and practices involving moral responsibility, things appear very much in a different light when considering the

[10] See also the pluralist suggestions of Watson (1996/2004), Macnamara (2011), and Mason (2019).

eye-poking (typical) adult. Here, we seem inclined to apply a further range of concepts and practices, above and beyond descriptions and explanations pertaining to causality, intentionality, or to management aimed at reducing similar things in the future. We might, for example, say that they were *wrong*, *bad*, or *mean* to poke the person in the eye. We might also demand to know why they poked the person in the eye, ask that they apologize for their behavior, and maybe even require them to accompany the person to the ER or compensate them for their inconvenience (if necessary).

So, while both scenarios involve human behavior with the same outcome, and while both are classifiable as unfortunate or harmful, they clearly generate very distinct intuitions in us regarding what response to the agent is appropriate. When an adult pokes someone in the eye, this triggers a readiness in others to view and interact with them in particular ways that are markedly different from when the same act is performed by a baby. This inclination seems to reflect an assumption that typical adult humans, but not babies, are eligible for assessments and responses of moral responsibility. This eligibility involves, at a first approximation, the attribution of moral agency.

When we ascribe moral responsibility or hold someone morally responsible, we thus seem to track and respond to features beyond particular behaviors or outcomes. We seem to track features that underly or constitute eligibility for ascriptions of moral responsibility. Such eligibility requires that the entity in question has more than a mere causal relationship to their actions and the consequences of these. They also need to be a *moral agent*. As such, moral agency is commonly assumed to be a prerequisite for ascriptions of moral responsibility for particular acts or outcomes.

According to a widely endorsed view, moral responsibility requires that the being in question stands in the right relation to their actions (or the outcome of their actions) (Talbert, 2022). This means that whatever features implicated as necessary by such conditions, underlie what it takes to be a moral agent. To understand the shape of debates regarding moral agency, its nature, and its presence or lack of presence in specific instances, it is therefore essential to grasp the basic conditions of moral responsibility. As the practice-focused approach to moral agency works by analyzing the conditions of participation in actual social practices of ascribing or withholding moral responsibility, this general structure of the relationship between the notions of moral agency and responsibility is further underlined. The next section provides a brief outline of two generally embraced basic conditions of moral responsibility.

2.1 Moral Responsibility: Basic Conditions, Positions and Debates

A common view among philosophers is that an agent needs to meet both an *epistemic* and a *control* condition to be morally responsible for something (see Rudy-Hiller, 2022a). These conditions can be traced back to Aristotle's excusing conditions of "force" and "ignorance" (Aristotle, ca 350 B. C. E./2002, 1109b30-1111b1; see also Fischer and Ravizza, 1998, Ch. 1; Nelkin & Rickless, 2017). The idea is that an agent needs to be free from undue force, coercion, compulsion, or other constraints, such that their choices and actions are *up to them*, and that they need to have certain knowledge or mental states, such as beliefs, about their action and its outcome, to be morally responsible for it.

It may be worth pointing out that distinguishing between a control and an epistemic condition does not necessarily mean that they are "entirely distinct, since how an agent controls her conduct will be in part a function of her epistemic resources" (McKenna, 2012, Ch. 1, p. 13). For example, if the adult eye-poker would lack, say, awareness of the position of their limbs (due to an injury), it seems reasonable to question whether they acted freely when they unknowingly poked someone in the eye.[11]

Nevertheless, the control and epistemic conditions serve as helpful starting points for thinking about, and analyzing, what different accounts say about the requirements and scope of moral agency. If control and knowledge are necessary for being morally responsible for an action or outcome, general *moral agency* requires whatever features that underly or enable such control and knowledge.

2.1.1 The Control Condition

The control condition of moral responsibility can be put in terms of having control over one's choices and behavior. A traditional understanding of what it means to have such control is that one has *free will* in the sense that one *could have done otherwise*. In other words, free will requires alternate possibilities that are available to the agent in the sense that the agent is able to effectively choose to act in different ways. Another central sense or interpretation of the control condition is *sourcehood*. This notion does not concern what one can do or could have done, but rather how one's actions were *actually brought about*. According to this notion, an agent has

[11] See McKenna (2012, Ch. 1, p. 13) for a helpful discussion.

control over their conduct if they are the ultimate causal source of their own actions (O'Connor & Franklin, 2022).

The assumption that free will or sourcehood (in the described senses) is necessary for moral agency has historically led much of the philosophical debate to focus on whether free will or sourcehood is compatible with living in a causally deterministic universe. Causal determinism means that any disposition, behavior, choice, or action that exists or occurs, however genuine, free, deliberate, or intentional it may seem, can be explained as caused by some prior event together with the laws of nature and the total past state of the universe. The question, therefore, is whether anyone can have control over their choices and behavior, in either of the described senses, if every event, including things like human disposition, deliberation and action, is causally necessitated in this way?

Free will incompatibilists have, in various ways, asserted that determinism is incompatible with free will. For instance, according to *the consequence argument*, if determinism is true, then no one seems to have power to alter the facts of the future (Ginet 1966, 1990; van Inwagen, 1975, 1983; Wiggins, 1973; Lamb, 1977; see also Vihvelin, 2022). Therefore, no person has freedom in the sense of being able to act differently than they in fact do. In a similar sense, the truth of determinism may appear to be at odds with control in terms of sourcehood. This is because, if determinism is true, then every fact about the past together with the laws of nature constitute sufficient conditions for every truth about the future (McKenna & Coates, 2021). Therefore, no person appears to be the ultimate causal source of their behavior.

If control requires the ability to do otherwise than one did (that is, free will) or to be the ultimate source of one's behavior, then control seems incompatible with existing in a causally deterministic world. This would mean that no entity, including typical adult humans, let alone babies, can have the kind of control required for moral responsibility. Consequently, on this line of reasoning, there would be no moral agents.[12] Hence, the consequence argument supports skepticism about free will and moral responsibility.

A classic rejection of such free will and responsibility skepticism is found in libertarian accounts of free will. Libertarians are incompatibilists. That is, like the free will skeptics, they maintain that causal determinism, if true, is incompatible with free will. However, they deny the truth of causal determinism, and believe that agents (sometimes) have free will. Libertarian accounts of free will internally

[12] Note, however, that this argument can still allow for independent justifications for holding people morally responsible.

disagree on two main questions: (i) the type of indeterminism required for free will and (ii) where in the process leading to an action indeterminism has to be located to result in free actions (Clarke et al., 2021).

According to non-causal libertarians, either free actions are not caused by anything, or such actions are non-deterministically caused. Free actions are "simple" or "basic" actions, such as "a mental event that does not consist of one mental event causing others". Such free actions also have an intrinsic *actish* "phenomenal quality" (Ginet, 1997, p. 89) or are intrinsically intentional (Ginet, 1989, 1990, 1997; Bergson, 1889/1910; McCann, 1998; see also Pereboom, 2014a).

Event-causal libertarianism instead holds that free actions are non-deterministically caused by precedent events. These are typically taken to be agent-involving events, such as beliefs, preferences, desires, or evaluative judgments. In addition, the production of action needs to involve "some type of indeterminacy" (Pereboom, 2014b, p. 30; see also Ekstrom, 2000, 2003; Mele, 1995, 1996, 2006; Kane, 1996).

Finally, according to agent-causal libertarians, agents nondeterministically cause free actions. This is since an agent, according to this type of libertarian theories, is itself a non-caused substance. As such, the actions caused by agents are distinct from other events in the world. When a marble rolls down a hill, it does so because of natural forces, its own structure, etc. But actions performed by agents cannot be explained by reference to prior states or any other state of events. This is because agents are uncaused causes of free decisions, with the power to start new causal chains (Kant, 1781/1787/1987; Berkeley, 1710/1998; Reid, 1788/1969; Chisholm, 1966; Griffith, 2010; see also Pereboom, 2014a).

A third position on the problem of causal determinism, besides skepticism and libertarianism, is compatibilism. According to compatibilism, free will, and consequently moral responsibility, is not incompatible with the truth of determinism (as held by both free will skeptics and libertarians). Thomas Hobbes expresses a classical strand of compatibilism when stating that a person acts freely as long as there is "no stop, in doing what he has the will, desire, or inclination to doe [*sic*]" (1651/1997, p.108). According to this account, freedom is the ability to act as one wants or desires, in the absence of any external impediments. Because determinism does not seem to entail that we never act the way we want or desire, or that we are always restricted or impeded, it is compatible with free will. This would mean that any agent who has a will or desires, and who can act in accordance with those, has free will.

However, this *one-way* (McKenna & Coates, 2021) interpretation of compatibilist free will has been criticized. Agents seem to be able to meet the stated conditions of willed actions and absence of external impediments, and still not appear to be truly free or morally responsible. For instance, people suffering from obsessive-compulsive disorder (OCD) or delusions can be said to act as they want and without finding any "stop" to do so. But if an agent pokes someone in the eye because they suffer from an obsessive compulsion to do so, or because they have the delusional belief that the person's eye is a light switch, their action does not seem to be free in any proper sense. This is, according to one line of incompatibilist critique, because freedom, and thus responsibility, requires that one could have acted otherwise than one did. Because the eye-poker suffering from OCD or a false fixed belief was unable to act in any other way than they did, they are not free, and hence not morally responsible for their action or its outcome.

An alternative compatibilist account, aimed at satisfying the notion of freedom as the ability to do otherwise, and that aims to remedy the mentioned issue posed by internal impediments, is found in the classical compatibilist *conditional analysis*. According to this strategy, an agent's ability to do otherwise is analyzed in conditional terms. An agent's ability to act differently than what they did is therefore determined by considering how they would have acted given a counterfactual will, desire, decision, etc. The compatibilist conditional analysis thus specifies the ability to do otherwise at the time of action as consisting in the following type of counterfactual truth: if the agent had wanted, decided, willed, or chosen, to do some action Y instead of X at that time, then they would have done Y and not X (Moore, 1912; Ayer 1954/2013).

But also this classical compatibilist notion of free will has been criticized from a control perspective. Assuming the truth of determinism, an agent is determined to have the wills, desires, or wants they have at the time of the action. As such, it does not matter to assert that they would have done differently given different wants. Because if determinism is true, all agents are determined to have the desires or wants they have, and so cannot act in any other way than they actually do. For example, the causal history of the OCD suffering eye-poker has led them to have a severe mental and behavioral disorder. They have an uncontrollable urge or need, to poke people in the eye. Hence, at the time of action, the eye-poker could not have had any other will, desire, or choice than they did. Despite acting in the absence of any external impediments or forces to do so, they could not have refrained from poking the person in the eye. As such, the conditional analysis fails to show that free will is compatible with determinism (van Inwagen 1983).

The philosophical discussion about whether free will (and moral responsibility) is (in)compatible with determinism continues to this day. The contemporary (in)compatibilist debate is, among other things, concerned with various forms of so-called manipulation cases and responses to such cases (Kane, 1996; Taylor, 1963; Pereboom, 2001; Mele, 1995; Mickelson, 2015, 2016). For the purposes of this thesis, however, I will now continue this introductory overview with relevant discussions in which the traditional questions about free will and determinism are (allegedly) not at issue.

2.1.2 The Epistemic Condition

As mentioned earlier, the second commonly assumed condition for moral responsibility is the epistemic, or knowledge, condition (Talbert, 2022). This condition is concerned with whether the agent had (or has) the cognitive state(s) required to be held morally responsible for a particular act. Were they sufficiently and properly aware of their choices or actions, the consequences and moral significance of those choices or actions, et cetera (Rudy-Hiller, 2022a)? For instance, did the baby and the adult know that they were about to poke someone in the eye? Were they aware that poking someone in the eye causes pain? And did they know that causing pain is, generally, morally bad or wrong?

The epistemic condition can be taken to imply that moral agency requires having features or capacities enabling *awareness of* one's action (what one is doing), the moral significance of one's action (in terms of its valence or its right- or wrong-, or good- or bad-making features), the consequences of one's actions, and, according to some, awareness of alternative courses of action. These are all examples of *contents* of awareness. But the epistemic condition can also be interpretated as requiring certain *kinds* of awareness. These can, for instance, be knowledge, reasonable or justified belief, beliefs simpliciter, et cetera. Another central issue regarding type of awareness is whether such awareness needs to be *occurrent* or merely *dispositional*. In other words, does one need to have subjectively phenomenally manifest states at the time of the action, or is it sufficient that one has implicit beliefs, such as dormant, unconscious, or tacit beliefs (Haji 1997; Peels 2011; Timpe 2011; Husak 2011; Nelkin & Rickless 2017; see Levy, 2014, Ch. 2)?[13]

While the control and epistemic conditions are typically assumed to be distinct, there are those who deny this. For example, some argue that the epistemic

[13] See Ramsey (2022) for a discussion about suggested distinctions between explicit and implicit mental representation.

condition is part of the control condition (Mele, 2010; Nelkin & Rickless, 2017), and some even claim that there is not any distinct epistemic condition to begin with (Björnsson, 2017; see also Rudy-Hiller, 2022b, note 1). In the last couple of decades however, there has been increasing attention to questions and challenges regarding moral responsibility allegedly raised specifically by the epistemic condition. Some of these challenges are even considered to constitute distinctive skeptical threats to the possibility of moral responsibility (Rudy-Hiller, 2022a).

Although it seems intuitive that moral responsibility requires awareness of certain kinds and with certain contents, the epistemic condition has been claimed to lead to a regress that may undermine ascriptions of moral responsibility in general. Consider the following example: an agent, such as our eye-poking adult, falsely believes that the victim is a mannequin. Since the agent is not aware that the person is, in fact, a living breathing human, they fail to satisfy the epistemic condition both with regards to awareness of action, and, in effect, also awareness of consequence and moral significance. Therefore, the eye-poker's ignorance seems to excuse them from blame.

However, it is widely assumed that ignorance can itself be blameworthy. Blameworthy ignorance is sometimes called *culpable ignorance* (H. M. Smith 1983). If the eye-poker is culpable for their ignorance of the fact that the victim is a human being, they may still be blameworthy for unwittingly causing the victim harm. But for the eye-poker's ignorance to be blameworthy they need to be blameworthy for the belief that the person was a mannequin or for lacking the correct belief (that the mannequin is in fact a person). What does it take to be blameworthy for beliefs or for lacking the correct ones? According to volitionists (Robichaud, 2014), ignorance is blameworthy if it is brought about by an action/omission which the agent had direct control over and which is generally wrong (Alston 1988; Zimmerman 2002; Rosen 2004; Levy, 2011). Such *benighting acts* (H. M. Smith, 1983) are thus acts or omissions which the agent brings about and the consequence of which is ignorance about the wrongness, badness, or the wrong- or bad-making features or facts, of one's behavior.

The eye-poker may, for instance, have performed benighting acts by failing to put on their glasses that morning, not asking a bystander to confirm the nature of the mannequin, and so on. But for the eye-poker to be blameworthy for any benighting act, they need to have performed them with the requisite awareness. If they lacked such awareness, however, they can only be blameworthy if it can be found that they performed another benighting act producing this lack of awareness. Establishing blameworthiness for acts or omissions for which the agent lacks

awareness, thus appears to run into a problem of regress. The only way to terminate the regress would seem to be to find a point in time where the eye-poker performed a so-called akratic act. That is, an act or omission performed in full awareness of its consequences and moral significance (its wrongness). This would mean that only akratic actions, or actions that result from an akratic act, can be blameworthy. This would, however, mean that most everyday ascriptions of blameworthiness may be unwarranted (Zimmerman, 1997; Levy, 2011; Rosen, 2004).

This revisionist upshot has, however, been challenged in various ways. For instance, some claim that clear-eyed akrasia (that is, having occurrent awareness of wrongdoing) is not necessary for blameworthiness. Non-occurent beliefs, such as dispositional or unconscious beliefs, are sufficient. Another way to stop the regress is to argue that agents can be blameworthy for their ignorance if their benighting acts are caused by epistemic vices, such as "overconfidence, arrogance, dismissiveness, laziness, dogmatism, incuriosity, self-indulgence, contempt, and so on" (W. J. FitzPatrick, 2008, p. 609).

Quality of will-theories of moral responsibility oppose the regress by linking blameworthiness to substandard care or regard. On this view, agents can be blameworthy for an action despite lacking awareness about the wrongness of the action because the action may still express lack of concern or regard for those affected (Fields 1994; Arpaly 2002, 2015; Harman 2011; Talbert 2013, 2017; Mason 2015; Björnsson 2017; Weatherson 2019, Ch. 5).

The most radical response to the regress argument is found in *capacitarianism* (Rudy-Hiller, 2017). According to proponents of this family of views, an agent can be blameworthy also for fully unwitting actions that are wrong. That is, blameworthiness does not even require factual knowledge. It is sufficient that the agent should and could have known better (W. J. FitzPatrick 2008; Sher 2009; Clarke, 2017; Rudy-Hiller 2017; Amaya & Doris 2015; Murray & Vargas, 2020).

In this thesis, the conditions of moral responsibility play central roles in all the papers. In Paper I, all aspects of both conditions are discussed at length. In Papers II and III the epistemic condition is in particular focus, while the control condition (primarily in terms of some sense of sourcehood) is discussed more indirectly, through the general discussion about what it means for someone to participate in a moral responsibility practice. In Paper IV the role of the conditions is less direct, but still present as the moral patiency of a being is there argued to morally warrant a particular perceptual stance in virtue of its epistemic value. Throughout, the arguments that are advanced assume the possibility of, or to some extent advances

specific variants of, a compatibilist moral responsibility approach that side-steps the traditional debates centering around determinism.

2.2 The Moral Responsibility Compatibilist Approach

A recent development in the moral agency and responsibility debate is to question the assumption that control in a sense that underlies the determinism-related debate is necessary for responsibility. Philosophers are increasingly asking questions about moral responsibility whose relevance do not bear on the truth of determinism. As we will see, many contemporary accounts ground moral agency and responsibility in features and practices beyond our control. Others, instead suggest less substantial or demanding conceptions of control or sourcehood, which are assumed to be compatible with determinism. These, *moral responsibility compatibilist* accounts of the conditions of moral responsibility make up the bulk of contemporary debates on moral agency and responsibility.

This section presents some influential moral responsibility compatibilist accounts. As we will see, various accounts assume, imply or actively defend different positions on what it means to be fit for ascriptions of moral responsibility. Some accounts incorporate both the control and epistemic condition (in some variation), while others only incorporate one, yet others neither.

2.2.1 Hierarchical Views

A moral responsibility compatibilist account that has been highly influential in contemporary discussions is Harry G. Frankfurt's (1971) hierarchical mesh theory. To fully appreciate Frankfurt's proposal, it is helpful to first take a look at his challenge to what he calls the Principle of Alternate Possibilities (PAP), which expresses a free will variant of the control condition (Frankfurt, 1969). Frankfurt uses a series of thought experiments to show that people can be morally responsible for their behavior despite not having been able to act differently. This is so, he argues, because there may be circumstances that render it impossible for a person to act differently, without those circumstances necessarily causing the person to act the way they did. Consider, for example, the following case:

> Suppose someone – Black, let us say – wants Jones$_4$ to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he

> prefers to avoid showing his hand unnecessarily. So he waits until $Jones_4$ is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that $Jones_4$ is going to decide to do something other than what he wants him to do. If it does become clear that $Jones_4$ is going to decide to do something else, Black takes effective steps to ensure that $Jones_4$ decides to do, and that he does do, what he wants him to do.[8] Whatever $Jones_4$'s initial preferences and inclinations, then, Black will have his way. (…) [But, n]ow suppose that Black never has to show his hand because $Jones_4$, for reasons of his own, decides to perform and does perform the very action Black wants him to perform. (Frankfurt, 1969, p. 835-6)

According to Frankfurt, cases like this show that agents can be morally responsible for their actions despite not having control in the sense of being able to act differently. Although Jones was not able to act in any other way than he did, his action was not *caused* by this fact. Jones acted the way he did for reasons unrelated to Black's readiness to manipulate him. He acted on reasons that were *his own*.

Following this rejection of PAP, Frankfurt proposes an identification model of sourcehood, according to which an agent can be said to be morally responsible for their actions if they themselves determine their actions. Identification accounts of sourcehood, draw a distinction between attitudes (like desires or motivations) that are internal to the agent and those that are external. Since agents can be properly identified with some of their attitudes, any actions that result from such internal desires and motivations are self-determined (also see Watson, 1975).[14]

Frankfurt's particular account explains moral responsibility using the expression of *free will* but using it in terms of the ability to act from a will that one endorses or identifies with (and thus unrelated to determinism). Frankfurt identifies an agent's will with one or more of their first-order desires, that is, desires about actions, such as eating chocolate, reading a book, or going for a jog. A will is, however, different from any desire in that it is an *effective* desire. That is, a will "moves (or will or would move) a person all the way to action" (1971, p. 8).

---

[14] Identification accounts of sourcehood have been proposed to be of roughly two kinds with regard to the nature of the identification relation between the agent and their attitude: authority accounts and authenticity accounts (Lippert-Rasmussen, 2003). Authority accounts identify an agent with "psychological elements that represent the person as the author" (2003, p. 368) of their own life. Authenticity accounts identify agents with attitudes that disclose their true self, such as "the person's deepest and most genuine commitments and desires" (2003, p. 368). Some, like Shoemaker (2015) combine these two accounts. In addition, the attitudes in question have been specified in terms of, for example, judgments or perceptions of the good (Watson 1975; Stump, 1988; Ekstrom 1993; Mitchell-Yellin 2015), loves or cares (Shoemaker 2003; Jaworska 2007; Sripada 2016), and higher order desires (Frankfurt, 1971) (see also O'Connor & Franklin, 2022).

Second-order desires are desires about desires and can take the form of either wanting to have a certain desire or of wanting a desire to be one's will. Frankfurt calls second-order desires of the latter kind second-order volitions. A person acts from a free will when their effective first-order desires align with their second-order volitions, and then they are morally responsible for this act and its consequences.

Frankfurt therefore claims that second-order volitions are "essential to being a person" (1971, p. 10) rather than a "wanton". Wantons, such as some animals as well as humans like young children and certain adults, have second-order desires, but lack second-order volitions. They may deliberate about how to attain the things they want, but they are never concerned with whether their desires *as such* are desirable. Consequently, a wanton does not reflect on what they should or should not desire, nor do they care which desire wins out in the end. As such, a wanton is not morally responsible. A person, on the other hand, through second-order volitions, identifies with some of their first-order desires rather than others, and cares about how they play out, and is therefore morally responsible.

To act freely, and subsequently be morally responsible for some action, Frankfurt holds, requires that one is moved by a will that one wants. Freedom of the will is therefore distinct from merely being free from external impediments or forces. A typical adult human can form second-order volitions and act from a free or unfree will. In contrast, babies cannot form second-order desires, and so they cannot be identified with any of their first-order desires. As such, the adult human eye-poker, assuming they were willing in the appropriate sense, but not the baby, is morally responsible for poking the person in the eye.

According to Frankfurt, some persons have an easier time exercising freedom of will than others. For instance, some "are naturally moved by kindness when they want to be kind, and by nastiness when they want to be nasty, without any explicit forethought and without any need for energetic self-control." (1971, p. 17). By contrast, others regularly struggle with exercising freedom of will. For instance, a person who suffers from OCD or addiction, and does not want this to be the case, is, in a sense, moved by forces other than their own. Having one's second-order volitions frustrated in this way renders the agent "a passive bystander" (1971, p. 17) to the will, and therefore not morally responsible. However, a *willing addict*, that is, a person who wants to have the first-order desires constitutive of addiction, and who acts on those desires, is morally responsible. Although their inclination to use the substance is over-determined by the first-order desire to take the drug,

the fact that they also want to be moved by this desire, renders the will to take drugs their own.

Frankfurt's hierarchical mesh theory has been argued to face several problems. One prominent line of criticism regards the "mesh problem". The way the mesh between an agent's second- and first-order desires comes about seems, contrary to Frankfurt's "willing addict" example, to be important. For example, the willing addict's second-order volition may very well have been caused by the drug itself. Weakened self-control is, after all, a hallmark of substance-addiction (see, for example, Heilig et al., 2021). In addition, an agent could have been manipulated (by hypnosis or advanced alien technology) to have the preferences they have.

A second challenge is found in the *hierarchical problem*. If it is the case that a person can be identified with the first-order will or wills that they endorse (via a second-order volition), the same appears to be true when considering conflicting second-order desires in relation to *third*-order ones which may align with the first-order desire. This problem of how to determine an agent's will seems to reappear at forever ascending levels (see, for example, Watson, 1975). Something more must be added to Frankfurt's account to render his weakened variant of the free will condition sufficient. Despite these problems, Frankfurt's arguments have inspired many compatibilist accounts where freedom or control are defined in other ways than in terms of alternate possibilities or where freedom does not constitute a requirement at all.

2.2.2 Consequentialist Views

An early type of moral responsibility compatibilist account of the latter kind can be found in consequentialist (or instrumentalist) accounts of moral responsibility. The main idea here is the claim that concepts and practices of moral agency and responsibility can be explained and by their potentially beneficial consequences. Despite the truth of determinism, reactions such as praise and blame, can influence people to behave better. When we call someone out on poking people in the eye, we are justified in doing so given that the perceived perpetrator will be less likely to do so in the future. Treating each other as moral agents is therefore an important means of encouraging or fostering favorable dispositions, choices, and actions, or of deterring people from harmful ones (Hobbes, 1654/1999). This means that moral agency simply requires whatever capacities or features necessary for being influenced or shaped in this particular way (Schlick, 1939/1966).

A well-known example of this forward-looking approach is found in Smart's (1961) moral influence account. Smart argued that when we blame someone for their behavior we are, apart from assessing it negatively also ascribing responsibility for the behavior. Such ascriptions merely consist in deeming the agent to be influenceable in the sense that they would have acted differently had they been given a motive to do so. Because of this, moral sanctions should be chosen with the aim of providing the target with incentives to refrain from harmful or wrongful behaviors.[15]

While Smart's account does provide a justification of our concepts and practices of holding responsible despite the truth of determinism, it has also been criticized on several grounds. A central criticism is that Smart's focus on beneficial outcomes disregards the conditions of moral responsibility as revealed in our practices of holding responsible. Efficacy or beneficial outcomes, the critics claim, play no central part in the reasons internal to our blaming responses. As such, the consequentialist account fails to accommodate features central to the phenomenology involved in ascriptions of moral agency and responsibility.

When we assess the eligibility and particular blameworthiness of other people, we are not doing so in a similar way to how, say, animal trainers assess the trainability of animals. When we, for instance, hold someone responsible for poking us in the eye, our judgment and our blaming reaction are not motivated by considerations about mere behavioral modification. Hence, Smart's original view does not properly distinguish between the psychology involved in, and conditions inherent to, ascriptions of moral responsibility and those involved in dressage or manipulation.

A related worry is that Smart's account fails to capture the centrality of desert to how many people view our moral responsibility practices. When we engage in discourses about the moral responsibility for some wrongful action, this is tied to a notion of assessing blameworthiness for this action, or to what extent its agent *deserves* to be blamed for it. But this aspect is completely ignored by a consequentialist account. Assuming that the concept of desert is essential, some claim that consequentialist account are not theories of moral responsibility in the first place (c.f. Vargas, 2022, p. 9). What is more, if the propriety of moral responsibility ascriptions and reactions depend on their potential to generate good behavior, we might very well have to disregard things like lack of control or ignorance when considering moral agency and responsibility. Smart's and other

[15] See also Nowell-Smith (1948).

consequentialist accounts might therefore imply that we have reason to hold innocent people, as well as intuitively exempted beings, responsible. This could, for instance, mean that it may be appropriate and justified to blame, and maybe even punish, babies insofar as doing so would, say, repel onlookers from wrongdoing.[16]

2.2.3 Reason-Based Views

Another influential moral responsibility compatibilist account can be found in a modification of Frankfurt's hierarchical mesh theory. According to proponents of the *Real Self View*,[17] the problem presented to us by compulsive agents, such as the willing addict, can be remedied by adjusting Frankfurt's original account. According to Watson (1975), the right type of relation an agent must have with her bodily movements to be morally responsible is the following: she has to be able not only to act on her volitions, but also to base her volitions on her *values*. As such, merely acting because we want something, even when second order endorsement of such wanting is present, is not sufficient. According to the Real Self View, then, the wanting addict is not morally responsible for her addiction. However, this is not because she lacks free will, but, rather, because her will is not based in her "evaluational system" (Watson, 1975, p. 220).

However, many object against Watson's suggested solution, and argue that the same problem that applies to the hierarchical mesh view also applies to the real-self view. That is, "all the problems that we have in accounting for the distinction between a good source and a bad source of our motivational make-up can be repeated with regard to the distinction between a good and a bad source of our valuational system" (Sie, 2005, p. 43). If we cannot account for the responsibility-undermining sources of motivation that explain and justify why the wanting addict is not responsible, while non-addicts generally are, this may be because we have indeterministic commitments.

A well-known type of example in the moral agency and responsibility literature assumed to illustrate this issue is the evil agent whose evil actions are due to an evaluational system caused by being brought up under *unfortunate* circumstances (see Chapter 4 for a more thorough discussion of this case). This has lead skeptics about sourcehood to claim that the requirements disclosed by our indeterministic

---

16 See Vargas (2022) for an overview of the various objections raised against consequentialist theories of moral agency.

17 Wolf (1990, Ch. 2) uses this label for this kind of view.

intuitions can be described in terms of "the necessity of alternative possibilities at the level of our intentions" (Sie, 2005, p. 44), the *Principle of Alternate Intentions*, or PAI for short. According to PAI, then, moral responsibility requires the ability to intend otherwise than one in fact did. What is missing in the willing addict case, according to the skeptics, is the psychological freedom to intend something other than the action one did intend.

An alternative compatibilist account that denies PAI is found in so-called *practical compatibilism* views of moral agency and responsibility (see Sie, 2005). This type of view recognizes the shortcomings of the Real Self View but deny that PAI undermines moral responsibility.

Wolf (1990), for instance, claims that agents are morally responsible for their actions if they are able "to do the right thing for the right reasons" (1990, p. 79). Being able to intend otherwise is, first and foremost, a practical question. People may not create themselves or act or intend in ways that show that they are free from the laws and circumstances of the world. However, Wolf argues, responsibility only requires freedom within the world, and that, she holds, is provided to us by reason. Because determinism does not seem to interfere with our ability to act in accordance with reason, PAI does not seem incompatible with determinism. Wolf thus commits herself to a view where physiological, but not psychological, determinism is true. According to Wolf there is nothing in the ways we see, react, and respond to each other that suggests that we lack the ability to intend to do otherwise. And for this reason, she argues, we do not have any strong reasons to doubt whether our moral responsibility practices are justified.

An additional variant of practical compatibilism is represented by Wallace's reason-based account (1994). In contrast to Wolf, Wallace refutes PAI. He argues that moral agency requires reflective self-control, that is, the ability to grasp and respond to moral reasons. An agent in possession of reflective self-control, and who fails to respond to a moral reason, is guilty of making a culpable choice. And we are justified in holding one another morally responsible to the extent that our actions express culpable choices.

The only way that conditions or circumstances can undermine moral responsibility, according to Wallace, is if they undermine or disable the very abilities underlying moral agency. Excuses excuse because they show that no moral wrong was made, and exemptions exempt because they show that the abilities required for moral agency were temporarily, or permanently, disabled. In this way, both Wolf's and Wallace's accounts defend moral responsibility compatibilism by reference to our ordinary moral responsibility practices and the attitudes and

assumptions displayed in these practices. They argue that an overall assumption in these practices is that most (typical adult) people possess abilities that enable them to know what to do and to act in accordance with this knowledge. This assumption, in turn, does not seem to be incompatible with determinism. Hence, the justifiability of our practices does not depend on the truth of determinism.

Another, influential, reason-based proposal is the *reason responsiveness view* defended by Fischer and Ravizza (1998). They claim that (even causally determined) agents can be morally responsible for their behavior in virtue of being *responsive to reasons*. According to this view, agents are morally responsible if they can be said to act from (their own) reasons. Moral agency, then, does require a capacity for control, albeit not in the traditional free will sense of having access to alternative possibilities.

The reason responsiveness view is based on a distinction between *regulative control* and *guidance control*. While the first involves "the power freely to do some act *A*, and the power freely to do something else instead" (Fischer & Ravizza, 1998, p. 31), the second type of control involves the capacity to act in response to reasons. Only control in this latter sense, guidance control, is required for moral responsibility on the reason responsiveness view. For an agent to have guidance control, they need to have a psychological reason-responsiveness *mechanism* that enables them to be receptive, and reactive, to rational considerations. What this means is that if an agent were to find themselves under different sufficient reasons, they would have acted differently. For this to hold, the reason responsiveness mechanism that moves them to act is, first, their *own*, and second, suitably responsive to reasons.

The mechanism is the agent's own in virtue of two features. First, they meet an *epistemic condition* of sorts, entailing that the agent "takes responsibility". This, in turn, involves having "a set of *dispositional* beliefs" (Fischer & Ravizza, 1998, p. 218) formed in a certain way (without force or manipulation). To take responsibility an agent needs to understand (dispositionally) that they are the source of their behavior, and that their actions affect the world around them.

Furthermore, the agent needs to recognize that their actions may be assessed by the moral community and consequently that it is *fair*, at least in certain circumstances, that they be subjected to responsibility assessments and ascriptions by others. Hence, when the agent is blamed or praised for their conduct, they acknowledge this by adopting a (corresponding) internal attitude, such as guilt, shame, or, if they disagree, indignation or self-pity. In addition to having this "cluster of beliefs" (Fischer & Ravizza, 1998, p. 217), the agent needs to have come

to adopt them in the appropriate way, that is, through moral education (or reflection later in life), without force or manipulation.

An agent's reason responsiveness mechanism is suitably responsive if it is moderately reason responsive, meaning that it is, first, "regularly" receptive to (moral reasons) and second, "at least weakly reactive to reasons" (Fischer & Ravizza, 1998, p. 244). For the mechanism to be regularly receptive to reasons, the agent must recognize reasons in such a way that their behavior would form an "*understandable pattern*" (1998, p. 71) from the perspective of a third-party observer (who has knowledge about the agent's beliefs and values). In addition, the agent's mechanism must (hypothetically) react to at least one sufficient reason to do otherwise (hypothetically) (see also Brink & Nelkin, 2013; Nelkin, 2011).

Following the reason responsiveness notion, the eye-poking adult would be morally responsible for their action insofar as they have their own mechanism that is suitably responsive to reasons. Most (typical) adults seem to meet the conditions for both ownership and suitable reason responsiveness. Therefore, typical adults are moral agents, and eligible for ascriptions and reactions of moral responsibility, according to the reason responsiveness view. Babies, on the other hand, do not seem to be able to take responsibility in the way outlined. It is also unclear whether babies can be said to be moderately reasons responsive. Therefore, babies cannot be considered morally responsible for poking people in the eyes, albeit small children may still be open to attempts at morally educating them (Fischer & Ravizza, 1998, p. 241; see also Chapter 4, sec. 4.1.6).

A prominent line of criticism against the reason responsiveness view is that it does not solve the problem posed by the *Source Incompatibilist Argument* (see, for example, Mele, 2019; Pereboom, 2001). According to this argument, free will requires that an agent is the ultimate source of one's behavior. However, determinism is incompatible with being an ultimate source of one's action. This fact remains, according to the source incompatibilist, equally for agents that are subject to, or free from, manipulation. Regardless of how one's reason responsiveness mechanism has developed, it will take the form it takes due to causes that are outside our control. Hence, there is no relevant difference between determinism and cases involving manipulation.

Recent years has seen numerous elaborations and variations of all of the responsibility compatibilist accounts described. For instance, contemporary moral influence accounts pair the reason responsiveness view with forward-looking elements of a consequentialist nature (Vargas, 2013; McGeer & Pettit, 2015; Jefferson, 2019). I will return to these modern, so-called, *instrumentalist* accounts in

Chapter 6, as I believe they offer an interesting alternative to, and have important implications for, the question of moral agency in non-paradigmatic entities. First, however, there is one last stop to make on the list of prominent moral responsibility compatibilist accounts.

2.3 Strawson and the Social Conception of Moral Responsibility

While different from each other, many contemporary moral responsibility compatibilist accounts, including some of those already mentioned, share the assumption that moral agency and moral responsibility are, in some sense, fundamentally social or inter-relational (Wolf, 1981; Watson, 1987/2004; Fischer & Ravizza, 1998; Russel, 2002; Wallace, 1994; Sie, 2005; McKenna, 2012; Shoemaker, 2017; Macnamara, 2015a; Mason, 2019; Vargas, 2013, McGeer, 2019; Hieronymi, 2020, see also McKenna & Coates, 2021).

The sociality premise has its root in P. F. Strawson's seminal paper "Freedom and Resentment" (1962/1982). According to Strawson, neither free will incompatibilists nor forward-looking compatibilists have got it right. This is because both camps have assumed a false understanding of moral responsibility. Hence, in order to consider and assess the force of any, be they *pessimist* or *optimist*, arguments about moral responsibility, Strawson suggests that we need to start from a proper understanding of this concept. Strawson's approach grounds our concepts and assumptions about moral agency and moral responsibility in a certain set of emotional reactions and attitudes, argued to be fundamental to human social life.

When we hold others morally responsible for some action, for instance, by blaming or praising them, we are responding emotionally "to the quality of others' wills towards us" (1962/1982, p. 70). As such, our moral responsibility practices are essentially emotional reactions reflecting a basic expectation of good will or concern for ourselves and others. Reactive attitudes, such as resentment, indignation, or guilt, are thus part of an inescapable human disposition to react to what we take to be expressions of substandard regard, such as attitudes of "contempt, indifference, or malevolence" (Strawson, 1962/1982, p. 63). As such, these reactions are fundamental to normal human relationships.

According to Strawson's account, then, how we hold each other responsible depends on perceptions about the attitudes of others toward us, rather than on assumptions about libertarian free will or how to influence behavior. In this way,

quality of will is assumed to be a basic condition of moral responsibility.[18] Moral agents are therefore beings to whom it is appropriate to take the *participant attitude* (or *stance* (Holton, 1994), or *standpoint* (Watson, 2014). From this perspective we see others as co-participants in normal human social life. As such, they are viewed as appropriate objects for the basic demand for due regard and thus candidates for reactive attitudes.

Hence, the compatibilist upshot seems to be that, not only are we unable to abandon our moral responsibility practices, but we do not appear to have any reasons pertaining to deterministic based skepticism to do so even if we could. This is because quality of will judgments appear to be fully compatible with determinism (Sie, 2005, Ch. 2; Wallace, 1994).[19]

In addition, Strawson believed that the various excusing pleas used in our moral responsibility practices provide additional support for the mentioned compatibilist claim. Some circumstances or factors, for instance like accidents or unforeseeable consequences, may show that the perceived perpetrator did not, in fact, harbor ill will or lack of concern. In such cases, we will typically suspend our reactive attitudes toward the action but retain the participant stance toward the agent.

An agent who, by mere accident (say by stumbling on a pebble) pokes us in the eye is not expressing any ill will and their action is therefore not an apt target for reactive attitudes of a blaming kind. The action, albeit harmful, did therefore not reflect a failure to meet the general interpersonal demand for regard or concern for the morally significant interests of others. It is not, then, lack of control by itself that explains why the agent is not blameworthy. Rather, lack of control is relevant to the extent that it separates the agent's will from the action or outcome. In other words, by showing why the action does or does not disclose any ill will. For instance, the adult eye-poker may be excused if they stumbled and accidentally poked the person in the eye (lack of control explains away ill will). They may also be excused if they, due to having very poor eyesight, thought that the victim was but a life-less mannequin, incapable of feeling pain (incorrect belief explains away ill will).

However, Strawson's account does not excuse on the basis of *moral* ignorance. That is, the adult eye-poker would not be excused if they, for instance, were unaware of the wrongness of causing other people pain. This is because such indifference or ignorance can, in fact, be seen as expressing substandard care or

---

[18] Note that some practice-focused accounts assume that quality of will replaces the traditional condition of control, knowledge, or both, while others treat it as an additional condition.

[19] See Szigeti (2012) for an argument against Strawson's "inescapability" (2012, p. 92) argument(s).

regard for others. The verdict of the potentially excusing factors in all these cases thus assume quality of will as the basic condition of our responsibility assessments (more on this in the following chapter, 3.2.2).

For some beings, however, we suspend the reactive attitudes not due to situational excusing conditions but for reasons pertaining to features of the very nature of the agents themselves. Strawson points to, among others, the following type of pleas: '"He's only a child', 'He's a hopeless schizophrenic', 'His mind has been systematically perverted', 'That's purely compulsive behaviour on his part"' (Strawson, 1962/1982, p. 65). Following Wallace's (1994) suggested distinction, these *exempting* pleas are distinct from mere *excuses* as they concern deficits or limitations in the psychological make-up of the exempted agent rather than excusing circumstances.

Small children and adults who are temporarily (say, because of drugs) or permanently (due to agency-undermining conditions or disabilities) psychologically *abnormal* are some examples of beings to whom we typically do not take the participant stance. Other examples include nonhuman animals and machines. Such human and nonhuman individuals are temporarily or permanently exempted from the practices surrounding our demand for due regard (Watson, 1987/2004).

According to Strawson, the reason for exemptions is, that it would simply not make sense to direct the usual demand for regard or concern towards certain agents. Such beings are not possible to engage with in ordinary human relationships, and therefore not eligible for the (whole range of) emotional responses characterizing normal social life. Instead, we often do, and *should*, approach such agents from an *objective stance*. From this perspective we view their behavior in purely causal terms.

A baby may, as mentioned, of course *cause* harm. However, their behavior can never be considered an appropriate object of resentment or indignation because they cannot be engaged with and related to in the normal way. Exempted agents may still be subjected to social policy, training, and management, but never responses that reflect expectations or express the demand for good will or due regard. To be fitting for this demand requires more than, say, awareness of consequences or morally relevant facts. For someone to be an apt target of reactive attitudes they also appear to need to possess some further competence or skillset enabling them to engage in moral responsibility practices.

In addition to the mentioned examples of exempted agents, the objective stance is available to us also for reasons beyond those of suitability for reactive

attitudes. Strawson argued that "we *have* this resource and can sometimes use it: as a refuge, say, from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity" (1962/1982, p. 67). However, such examples should be seen as limited exceptions, akin to excusing conditions. We cannot be motivated nor forced to take the objective stance toward typical adult humans as our normal stance by "a general theoretical conviction" such as the truth of determinism (1962/1982, p. 68) nor because of reasons of convenience or "the efficacy of these practices in regulating behavior" (1962/1982, p. 61).

This is because moral reactions simply do not spring from assessments about freedom or desirable outcomes. Instead, they express concerns and demands about the attitudes of other people toward us and others. Hence, while a small child may poke us in the eye, their action does not express substandard regard (in the way relevant for reactive attitudes). In this way, our withholding blame or praise in light of certain deficits or abnormalities is claimed to be explained by the quality of will analysis, rather than by appeal to things like control, and in turn support that analysis. In this way, excusing and exempting reasons cannot be extrapolated into a thesis that reflects the thesis of determinism (Wallace, 1994, sec. 5.4). A general objective stance is unavailable to us because seeing and treating others as apt objects of reactive attitudes is an innate human disposition, central to normal social life, and thus unavoidable.[20]

Strawson's moral responsibility compatibilist argument has, however, been criticized. For instance, some claim that the truth of determinism can imply that even typical adult humans are deficient or abnormal in ways that might call into question their openness to reactive attitudes. For example, Russel argues that an implication of Strawson's rationalistic argument is that ["i]f the thesis of determinism is true … then we are, indeed, all morally incapacitated." (1992, p. 296) and subsequently have to extend the objective stance to everyone. Others question Strawson's claim that we are psychologically unable to refrain from our responsibility practices, or his claim that abandoning them would necessarily mean abandoning normal social relationships (Nelkin, 2011, Ch. 2; G. Strawson, 2010, Ch. 5; Watson 1987/2004, pp. 255–258; Sommers, 2007).

---

[20] Russel (1992), Wallace (1994) and McKenna (2005), among others, claim that Strawson's account implies additional arguments beside his naturalistic one. Also see Waller (2006), Pereboom (2014a, Ch. 8), and Hieronymi (2020) for suggestions of the compatibilist arguments provided by Strawson. See J. Campbell (2017) for an overview and analysis of the possible compatibilist arguments in Strawson's account.

Another important objection to Strawson's account is its lack of independent responsibility facts. As the account stands, the propriety of holding responsible does not depend on a particular view of moral responsibility, on which x is morally responsible under conditions y and z. Instead, according to Strawson, someone is morally responsible in virtue of being subjected to reactive attitudes. However, the practices of holding responsible, as described by Strawson, do not seem to provide any practice-independent responsibility facts as reactive attitudes do not involve or express propositions (Watson 1987/2004, p. 222). [21]

Despite these criticisms, Strawson's theory remains an important influence on much of the moral responsibility compatibilist work that has followed (McKenna, 2005). One apparent upshot of his social conception of moral responsibility is that the truth of determinism, and its implication for free will, are beside the point (Strawson, 1962/1982). We are innately disposed to react in certain ways to behavior exceeding or falling short of our expectation for due regard. Our propensity toward the reactive attitudes is an inescapable part of our human nature and can therefore not be replaced by a general objective stance.

In addition, our practices of moral responsibility track and respond to standards internal to these practices. As such, they do not depend on any external facts, such as theoretical convictions regarding the truth of determinism.[22] Strawson writes that his aim has been "to represent skeptical arguments and rational counter-arguments as equally idle—not senseless, but idle—since what we have here are original, natural, inescapable commitments which we neither choose nor could give up" (Strawson, 1985, p. 28).

---

[21] Note that the assumed nature of the relationship between holding and being morally responsible, as well as the question of which is metaphysically more basic, are answered differently by contemporary Strawsonian accounts. For instance, some take it that being responsible is more fundamental (Brink & Nelkin, 2013). According to others, however, a thorough Strawsonian account of moral responsibility gives that being responsible is a function of holding responsible rather than the other way around (Shoemaker, 2017). A third position is defended by McKenna, among others, who claims that "neither being nor holding morally responsible can be regarded as metaphysically more basic than the other, and that each is significantly implicated in a direct metaphysical explanation of the other." (2012, p. 81). See De Mesel (2022) for an overview of positions in this debate as well as a suggested solution.

[22] McKenna calls these two arguments the "Psychological Impossibility Argument" (2005, p. 166) and the "Internal Justification Argument" (2005, p. 167). Pereboom (2001) and Watson (2014) make similar distinctions between a psychological and a normative argument. See Szigeti for a critical examination of what he suggests are "at least four inescapability arguments" (2012, p. 94) in Strawson's account. See Sars for a suggestion that Strawson's argument reveals both an "Incapacity Argument" (2022, p. 78) and an "Inconceivability Argument" (2022, p. 82), and that the latter, along with its alleged implications, have been largely overlooked by both adherents and critics of Strawson.

Looking at, and starting from, moral responsibility practices thus serves to provide support for moral responsibility compatibilism (McKenna, 2005). In addition, Strawson's naturalistic strategy points toward a possible, underappreciated, direction for questions about nonhuman and nonstandard moral agency. Namely, that "the only way we'll be able to determine the capacities required for morally responsible agency is by examining our practices, including our susceptibility to the moral emotions and the ways we tend to respond to wrongdoing" (Tognazzini, 2015, p. 20).[23] This brings us to the methodology and basic premise of this thesis, namely, the practice-focused approach to moral agency. This approach is further specified and developed in the next chapter. So too are my reasons for preferring it to a capacity-focused approach as well as my worries about, and contributions to, this type of approach.

[23] It is worth noting that Strawson does not seem to explicitly make this point himself.

3 Why the Practice-Focused Approach

In this chapter I explain what the practice-focused approach is about, motivate why I use this approach, clarify how I use it, and account for what I take the implications of this approach to be. The chapter begins with motivating and specifying the practice-focused approach. This is then followed by an account of prominent views which incorporate central features of Strawson's original conception (introduced in the preceding chapter), and which, to various degrees, can be characterized as practice-focused. This is followed by a discussion about the possible virtues of the practice-focused approach in general, as well as for investigating moral agency in nonstandard cases in particular. The chapter is concluded with a suggestion for two desiderata for a practice-focused account of moral agency and how these desiderata have informed each of the four papers.

3.1 Centering the Practice

This thesis has been developed within a theoretical framework of Strawson's account of moral responsibility in general, and what I call the practice-focused approach to moral agency in particular. This approach is set out in Paper II and further developed in Papers III and IV. Its main idea is that the nature and requirements of moral agency are determined by the nature and requirements of participating in moral responsibility practices.

3.1.1 The Capacity-Focused Approach

My path into the practice-focused approach is very much due to the flaws and weaknesses I have observed in the traditionally dominant idea of the philosophy of moral agency, what I here call the capacity-focused approach. This idea, simply put, assumes that the moral agency of a being is determined by its possession of certain theoretically pre-defined intraindividual capacities.

My move towards the practice-focused approach was, in part, originally motivated by the challenges in assuming the capacity-focused approach identified in

Paper I. The main thesis of that paper is that the artificial moral agency (AMA) debate has been overly focused on apparent disagreements regarding theoretical issues relating to determinism and free will, as well as questions relating to the importance of (phenomenal) consciousness, and other (advanced) cognitive capacities assumed to be essential for moral agency.

However, disagreements about these theoretical issues seem to be of little use for solving the practical concerns and issues related to the possibility of artificial moral agency. A possible solution to this standstill would therefore be to ask how we *should* approach these questions to begin with. Instead of the widely assumed "theory-first" approach prevalent in the capacity-focused debate, we suggest that one asks to what extent, and how, machines *should* be included in contexts and practices where moral agency is normally assumed.

The full step towards a practice- rather than capacity-focused approach resulted when I turned my gaze toward discussions about moral agency in nonhuman animals. In Paper II, I claim that like the artificial moral agency debate, discussions about animal moral agency often involve disagreements about the relevance and validity of certain (standard) *intraindividual* features or properties, such as metacognitive capacities, like reflection, evaluation, and self-consciousness.

For instance, skeptics regarding animal moral agency, like Korsgaard (2006; 2010), Ayala (2010), Musschenga (2015), and Kitcher (2011), typically defend their negative verdict on the assumption that metacognitive capacities enable or constitute requisite knowledge and control for moral agency, while lacking such capacities provides reasons for exempting someone from moral agency.

Interestingly, however, a surprising number of proponents of animal moral agency likewise maintain that nonhuman animals may only be ascribed less robust types of moral agency due to their lack of certain internally construed features, or properties such as capacities for conscious reflection or deliberate moral reasoning (Rowlands, 2012; Shapiro, 2006; Sapontzis, 1980, 1987; de Waal, 2006). Hence, although animals can show seemingly virtuous behaviors, dispositions, or expressive emotions, like compassion, selflessness, or altruism, none of these are generally deemed sufficient for the ascription of moral responsibility/moral agency, even according to the more permissive views of animal moral agency (see Chapter 4, sec. 4.2.2).

The capacity-focused approach prevalent in the animal moral agency discussion thus similarly assumes a theory- or concept-first way of thinking about moral agency. It starts from the assumption that certain predefined (intraindividual) features or properties, such as rational deliberation, are necessary

for moral agency, and then goes on to consider the possibility of such features in other animals. This theoretical strategy has also been called the *rationalistic strategy* to moral responsibility (Russel, 2002), and assumes the following method:

> (1) a coherent and intelligible concept of responsibility and an account of its conditions of applicability; (2) that we show that this concept does indeed have some application (to human beings); *and* (3) that we tackle this problem in this order—from the concept to its application. (Russel, 2002, p. 172).[24]

3.1.2 The Naturalistic Strategy

An alternative approach is found in the *naturalistic* strategy, fundamental to Strawson's conception (Strawson, 1962/1982, 1985 Ch. 1; Russel, 1992; Wallace, 1994; McKenna, 2005) as well as David Hume's (1739-40/1978, 1757/1875, 1777/1975) theory of moral responsibility (Russel, 2002). A central premise of this strategy is that a good theory about moral responsibility and agency needs to account for, and recognize, the everyday reality of the actual use and application of these concepts (see Hieronymi, 2020; J. Campbell, 2017; Russel, 1992, 2002). To do this, Strawson urges us:

> to try to keep before our minds something it is easy to forget when we are engaged in philosophy, especially in our cool, contemporary style, viz. what it is actually like to be involved in ordinary inter-personal relationships" (Strawson, 1962/1982, p. 64).

Similarly, on a naturalistic reading of Hume, Russel (2002) believes that we must:

> eschew rationalistic, a priori investigations into the nature and conditions of responsibility in favour of a more empirical approach. More specifically, we must carefully examine and describe the attitudes, sentiments, and practices associated with responsibility *as we find them*. Only then will we be in a position to effectively criticize and evaluate the rationality of the attitudes, sentiments, and practices in question. (Russel, 2002, p. 173).

In recent years, similar ideas have been employed in philosophical discussions about moral psychology and normative cognition and behavior more broadly (Westra & Andrews, 2022; Heyes, 2023). The common method both within philosophy and empirical science has been to conceptualize and identify a domain-specific psychological capacity, assumed to underpin the "capacity to acquire,

---

[24] See also Vargas (2022) for a comparison between "phenomenalist" and "conceptualist" (2022, p. 9) approaches, and Argetsinger and Vargas (2022) for a comparison between "concept-first" and "practice-first" methodologies (2022, p. 47).

enforce, and comply with the norms of one's community" (Westra & Andrews, 2022, p. 6; see also Heyes, 2023; Sripada & Stich, 2007). Such theorizing often proceeds from paradigmatic, "clear-cut-cases" (Westra & Andrews, 2022, p. 8), involving linguistic expressions of, for example, moral judgments.

However, critics argue that the upshot of this theory-first, atomistic, and psychologically focused, approach, has been that behaviors that do not seem to be driven by one's favored predetermined ought-thoughts are dismissed as irrelevant and not the real deal. By consequence, accounts that assume a theory-first approach, run the risk of merely reflecting the "most cognitively advanced, institutionalized forms" of the target behavior. Westra and Andrews (2022), among others,[25] therefore worry that this approach may set the bar too high and "not provide a reliable guide to understanding how social norms develop, how they have evolved, or how they manifest themselves across different social environments." (Westra & Andrews, 2022, p. 8).

This is claimed to motivate a shift from the theory-first, "inside-out", strategy to normative cognition. Instead of "a priori conceptual analyses" or "armchair intuitions" the question of the nature and underpinnings of social norms should be approached from the "outside-in" (Westra & Andrews, 2022, pp. 8-9). This means shifting focus from predetermined ought-thoughts toward "readily observed and measurable attributes" (Westra & Andrews, 2022, p. 25), like "normative regularities and the patterns of social interaction that constitute them" (Westra & Andrews, 2022, p. 9).

Among the type of social normative behaviors and cognitions targeted by this methodological shift of moral psychology, "social maintenance" behaviors (Westra & Andrews, 2022, p. 10), such as blaming and praising, have been lifted as a particularly fruitful avenue for inferring the existence of normative prescriptions or prohibitions, even when these normative expectations are not "explicitly avowed" (Westra & Andrews, 2022, footnote 5, p. 11). As we will see, the methodological emphasis on such behaviors is fundamental also for practice-focused accounts of moral agency and responsibility.

The practice-first strategy is therefore the view that moral responsibility practices, such as those of holding responsible, "are *epistemically prior* to being responsible" (De Mesel, 2022, p. 1902).[26] These practices offer the crucial starting

---

[25] See also Buckner (2013).

[26] Note that I do not assume, nor take a stand on, the further claim that our practices (of, say, holding others responsible) are *metaphysically* prior to being responsible (see also Shoemaker, 2017, McKenna, 2012, and De Mesel, 2022).

point for questions about the nature and conditions of moral responsibility. And, by consequence, for determining whether ascribing moral agency to particularly controversial entities is warranted. The practice-focused approach likewise allows us to assess the relevance of any alleged requirement of moral agency. If, for example, certain traditional requirements, such as free will, or rational deliberation are not present, or of marginal importance to, the actual discourses, decisions, behaviors, attitudes, or inclinations found in (human) moral responsibility practices, we seem to have reason to question and re-evaluate their relevance as requirements for moral agency. By extension, we would also have reason to question any claims and arguments derived from such requirements for or against the possibility of nonhuman, and other nonstandard, moral agents.

For these reasons, in this thesis, I investigate the possibility of moral agency in a specific domain of typically exempted nonhuman cases considering a practice-focused approach to moral agency. Instead of assuming from the get-go that moral responsibility requires "some specific property, power, or quality" (Russel, 2002, p. 171), such as free will or rational deliberation, and then examining the relevance and validity of this concept, the alternative, practice-focused, strategy suggests that we should start with looking at the actual practices where moral agency and responsibility are assumed. This strategy emphasizes the importance of understanding how moral responsibility figures in real-life settings. How do these interpersonal attitudes and behavioral patterns look, and what do they imply for the question of what it means, in practice, to be a moral agent (McKenna 2012, Ch. 1; McGeer, 2019)?

While Strawson's account clearly emphasizes what in norm psychology is commonly referred to as *social maintenance* behaviors as the target behavior, he is much less clear on the implications one can draw from these, in terms of capacities, skills or requirements of moral agency (see Watson, 1987/2004 and Russel, 2004). Thankfully, there is a rich literature inspired by Strawson's conception aimed at, among other things, providing exactly this. The following section presents some prominent ideas and developments in this literature in order to further flesh out the details of my own practice-focused approach to moral agency.

3.2 Prominent Practice-Focused Accounts

This section exemplifies some prominent ideas and themes of contemporary accounts that to various degrees follow Strawson's naturalistic lead and which may be said to fit the practice-focused approach to moral agency. Moral agency is

understood in terms of what is required for participation or engagement in moral responsibility practice. It does so by looking at and determining the nature of the actual practices, then using this analysis to arrive at the requirements of moral agency. As we will see, the accounts below second Strawson's view that moral agency should be understood as eligibility for the basic demand or expectation of due regard as expressed in ascriptions of moral responsibility. However, they diverge on what this entails and whether, and, if so, what other features might be required.

3.2.1 Communicative Views

An influential account of Strawson's notion of reactive attitudes is found in Watson's claim that such attitudes are "forms of communication" (Watson, 1987/2004, p. 30). This characterization of the reactive attitudes is argued to explain and justify why people tend to exempt many agents from moral responsibility, despite the fact that those agents seem to express ill will or lack of concern. Watson points out that "[a] child can be malicious, a psychotic can be hostile, a sociopath indifferent, a person under great strain can be rude, a woman or man "unfortunate in formative circumstances" can be cruel (Watson, 1987/2004, p. 28). To make sense of why it is still appropriate to exempt the mentioned agents, Watson believes, a communicative interpretation of Strawson's theory is called for. He argues that when we, for instance, feel and express resentment, we are basically *addressing* the target of this attitude to communicate the basic demand for reasonable regard.

An implication of the suggested communicative understanding is that since reactive attitudes involve a 'moral address' toward someone, the content of this address must be "intelligible" to the recipient. Therefore, reactive attitudes and are not suitable in the case of certain beings who lack the type of moral understanding required to appreciate the address. Watson hence argues that we can make sense of exemptions by recognizing the communicative nature of accountability responses. Following the communicative account, exempting, say, small children can be justified because it would be unreasonable to subject someone to a message that they cannot fully grasp.

The communicative account of reactive attitudes is widely assumed among supporters of the Strawsonian naturalistic strategy for moral responsibility (Wallace, 1994; Shoemaker, 2007, 2013, 2015; Macnamara, 2015a; Darwall, 2006; McGeer, 2012, 2013; Mason, 2019, Vargas, 2013; Sie, 2005; Fricker, 2016). A

particularly developed such account is found in McKenna's (1998, 2012) conversational theory which plays an important role in Paper II. According to Mckenna, a moral agent can be compared to a "competent speaker of a natural language" who has "skills both to express herself, thereby making contributions to dialogue, and also the interpretive skills needed to understand others." (2012, p. 85).

Analogously, McKenna claims that a moral agent's "acting skills and her holding-responsible skills are similarly enmeshed" (2012, p. 86). To understand and anticipate how others will come to interpret one's own actions, a moral agent needs to understand and be able to interpret the (moral) significance of the actions of others. In addition, a moral agent must understand the specific reactions characteristic of our practices of holding responsible.[27] Therefore, holding someone responsible is appropriate to the extent that it constitutes "meaningful, fitting or intelligible conversational response" (McKenna, 2012, p. 90). Some beings are exempt from moral responsibility because their behavior just cannot "reflect the sort of moral quality that a person with appropriate moral understanding and imagination could be taken to intend" (McKenna, 2005, p. 172).

I am very sympathetic to the communicative view of moral responsibility, and the arguments developed in Paper III are largely based on such an understanding of moral responsibility practices. Contrary to McKenna, however, I think that it is meaningful to distinguish between the features required for recognizing and responding to the morally relevant features of a situation, and the features required for engaging, and being engaged with, in moral exchanges. While these features often coincide, I argue that there are social contexts where they come apart in significant ways (Paper III).

3.2.2 Moral Competence

One implication of the communicative view is that eligibility for reactive attitudes requires something beyond mere (practical) rationality. This brings us to the next prominent theme in contemporary practice-focused views – the view that moral agency requires some kind of *moral* competence. According to Wolf, moral agency cannot merely require the capacity to act in accordance with one's values and commitments. To be eligible for ascriptions of moral responsibility in a stronger,

---

[27] In contrast to natural dialogue, however, McKenna claims that moral conversations are not initiated upon the expression of reactive attitudes. Instead, reactive attitudes are responses to the message implicit or implied in the behavior of the transgressor. When we act, our conduct can be said to *invite* moral conversations in virtue of revealing our quality of will. See Paper III, sec. 2.2 for a description of one type of moral exchange trajectory.

more real, sense, an agent also needs to be "sane" (Wolf, 1987/2013, p. 335) in the sense of having "normative competence" (Wolf, 1990, p. 129). This is the capacity to acquire the *right* values.[28] Or, in other words, the ability to recognize and respond to *moral* considerations.

Similarly, Russel thinks that for an entity to be a "full participant in the moral community", they need to have something beyond reason responsiveness or rational self-control (Russel, 2004, p. 300). Moral agency requires what Russel calls a "moral sense", that is, the ability to "feel and understand moral sentiments or reactive attitudes" (Russel, 2004, p. 293). An agent who lacks moral sense, lacks the "'internal' system of sanctions (or incentives) as associated with moral sentiments" (Russell, 2004, p. 296). This has real-life implications, as the agent's ability to identify and direct themselves by moral considerations will be impaired. To support his point, Russel asks us to imagine a person, Jill, who lacks the ability to feel or express fear. While Jill can learn about and know conditions that are dangerous or potentially harmful, this "external" or "superficial" understanding can never enable her to feel fear or be moved by it the way it motivates others (Russel, 2004, pp. 294-5). In this way, Russel argues, moral sense is a basic requirement of moral agency, and it would be "both unreasonable and unfair to communicate and reason with" someone who lacks it (2004, p. 295).

Considerations about fairness have been more saliently invoked to motivate a moral competence requirement for moral agency. Wallace states that while "[a]nimals and young children may be agents" in a "minimal sense" (Wallace, 1994, p. 13) of being able to act in accordance with the goals of their desires, moral agency requires that the other party has fallen short of our expectation in virtue of a *culpable choice*. Our practices thus reflect a commitment to principles of fairness that govern when holding someone responsible can be justified. Because moral agency denotes eligibility for moral sanctioning, Wallace believes that it requires a "normative competence in virtue of which one is able to grasp moral reasons and to control one's behavior by their light" (Wallace, 1994, p. 15).

The specific features or properties required for moral competence, are therefore often suggested in terms of, at least in part, "internalistically construed" (Sneddon, 2005, p. 241), features of individual agents, such as moral sanity, namely, "the ability to form her values on the basis of what is True and Good" (Wolf, 1990, p. 75), the powers of reflective self-control (Wallace, 1994), an evaluational

[28] I will return to Wolf's argument, and her famous example of JoJo, in the next chapter (sec. 4.1.2).

system (Watson, 1975), a moral sense (Russel, 2004), or the capacities for quality of will (Shoemaker, 2013).

However, while moral competence *can* be understood in terms such criteria, there is reason to question an entirely intra-individualistic conception. For one, intra-individually construed psychological criteria "fail to capture the fact that moral responsibility presupposes the possibility of a distinctive sort of interaction with those holding morally responsible" (McKenna, 2012, p. 80). Assuming a practice-focused approach, moral agency is realized and thus demonstrated socially, as opposed to individually (Sneddon, 2005). Hence, such "*a priori* theorizing about the universal psychological conditions of moral responsibility" need to be replaced by "*an posteriori* and locally contingent approach" (Sneddon, 2005, p. 261).

This, in turn, raises an additional possible issue. Intra-individual requirements may pose a challenge to the possibility of conducting comparative assessments of moral agency. The presence or absence of required features needs to be assessable in practice. Hence, any requirements need to be operationalizable in a way that is not only "*essentially* interpersonal" (McKenna, 2013, p. 128) but also measurable for other participants of a moral responsibility practice so that they can detect when someone has the required moral competence and when they have not.

The need for more readily observable features of moral competence is, of course, also supported by the general notion of a communicative account of moral responsibility. For example, the idea that the agent has some inclination for and understanding of "emotional communication" (Shoemaker, 2013, p. 118), the language of moral responsibility practices (McKenna, 2012), or a "capacity to secure uptake of the *to be communicated* emotional responses of anger and gratitude" (Shoemaker, 2013, p. 118). In Paper III, this aspect is discussed in terms of the competence required to engage others and be engaged with in *moral exchanges* (See also McKenna, 2012; McGeer, 2012).

3.2.3 Different *Faces* of Moral Responsibility

Another influential practice-focused view is the idea that our moral responsibility practices involve or make explicit distinct dimensions or forms of moral responsibility. Attending to these distinct *faces* of moral responsibility can serve to dissolve certain disagreements about moral responsibility as well as explain our

ambivalence toward certain agents, such as psychopaths and small children (Watson, 1996/2004).[29]

This type of pluralistic account of moral agency and responsibility originated in response to Wolf's assertion that moral agency is an all or nothing affair where the requirements of viewing someone as a moral agent have to go beyond "identifying her particularly crucial role in the causal series that brings about the event in question" (Wolf, 1990, p. 40). Rather, blame and praise are responses within a social practice in which we make and direct demands about conduct. Watson questioned this assumption by arguing that moral responsibility involves at least two faces. The first, aretaic or *attributability face*, of responsibility is shown in character judgments such as "she is untrustworthy", or "he is very generous". Such evaluations are sensitive to the values and commitments of a person ("as an adopter of ends"), and therefore requires that the agent has "the capacity to conform her desires and conduct to her deepest values" (Watson, 1996/2004, p. 261) and "the intelligence and sensibility to comprehend at least the normative concepts in terms of which the relevant forms of appraisal are conceived" (1996/2004, p. 282). The second, *accountability face*, of responsibility is implicated when we *hold* one another to account. Such responses are sensitive to "the faults identified in aretaic blame" (1996/2004, p. 278) but imply a further kind of communicative competence or knowledge as well, typically assumed to be implied by the communicative account of moral responsibility practices.

A particularly developed and influential pluralistic responsibility account is found in Shoemaker's (2007, 2013, 2015) tripartite theory.[30] Shoemaker argues that there are *three* distinct types of responsibility emotions responding to three distinct qualities of will. Agents who are eligible for moral responsibility can be eligible in one or several of the following ways: in virtue of their character, and thus eligible for aretaic praise or blame, in virtue of the judgments they make, and thus be eligible for demands to offer reasons or explanations, and, in virtue of their regard for others, and thus eligible for being held accountable via moral demands for acknowledgment (Shoemaker, 2013, 2015).

I do not want to take a definitive stand regarding pluralism or monism about moral responsibility (and agency), nor defend or challenge any particular pluralist account. What I will say, however, is that the primary interest of this dissertation concerns moral agency in terms of a socially situated competence and

[29] The next chapter (4) provides an extensive account of various philosophical discussions concerning the moral agency of, among others, nonstandard human cases.

[30] Other pluralistic suggestions have been made by Macnamara (2011) and Mason (2019).

corresponding eligibility. That is, moral agency is understood in a sense that most closely follows the inherently social faces of moral responsibility, akin to answerability and accountability. At the same time, I concede that pluralistic suggestions do seem able to explain several diverging intuitions, an idea harnessed to deal with some of the conceptual confusion discussed in Paper I. For instance, various positions about the content and target of moral responsibility (ascriptions) may be made more precise by employing a pluralistic account, and using such an account may help distinguishing more clearly between various levels or forms of moral agency. I will therefore utilize and apply pluralistic terminology in the next chapter to categorize and clarify various positions in discussions about moral agency in nonstandard cases.

3.2.4 Methodological Challenges

While there are differences between all the accounts presented in this chapter, their shared practice-focused nature implies a number of commonalities. They share, for instance, a methodology whereby certain practices, namely those pertaining to the *social maintenance* aspect of morality (such as the enforcement of expectations, norms, or standards), provide the data of choice for considerations about the nature, conditions, and requirements of moral agency. Moral agency is assumed to require features that make an agent eligible for the various reactions and treatments of moral responsibility in these practices. Our inclination for reactions of, say, praise and blame are then taken to express or reveal a commonplace demand or expectation for sufficient concern or regard. Hence, being eligible for participation in moral responsibility practices requires that the agent is an appropriate object or target of said basic demand or expectation, as expressed through the reactive attitudes.

Despite this general methodological approach, there is scant guidance on what is to be a proper data selection and interpretation. For instance, it is not apparent *how* one ought to gain knowledge about the crucial details about these practices. Are introspection, participant recollection, and thought experiments sufficient instruments? Or should these *armchair* methods be complemented with, or even replaced by, systematic, nonparticipant, observation through scientific empirical studies of real-life social morality?

Furthermore, it is not given what the target attitudes and behaviors are. What more exact aspects of (human) social interaction should be considered part of moral responsibility practices, and which should not? Likewise, opinions differ on

the content of moral appraisal. What kind of moral evaluation is blame – is its nature deontic, axiological, aretaic or something else (McKenna, 2013)?[31] Some authors propose a narrow conception of the range and content of reactions and ascriptions of moral responsibility (Wallace, 1994), while others defend a wider one (McKenna, 2013; Macnamara, 2013; Holroyd, 2018).

Likewise, it is not clear how one should derive normative criteria (in terms of, say, fittingness or fairness) from a mere descriptive account of the practices as they stand. Surely, not all instances of holding responsible reflect or reveal appropriate or fitting conditions and requirements of moral responsibility and agency? Likewise, the actual inclusion or exclusion of agents from such practices cannot, in every case, be legitimate? The question remains how one should bridge this gap between what we see in the reality of moral communicative practice and what of that *should* be part of that practice.

In light of these methodological questions, the next section distinguishes two possible objectives of adopting a practice-focused approach: what I take to be the traditional Strawsonian motivation of providing a naturalistic compatibilist defense of moral responsibility *and* the central objective of this thesis – namely, to investigate the possibility of moral agency in nonstandard, especially nonhuman, cases. I argue that, appearances notwithstanding, a central criterion of both of the mentioned objectives is that the account in question starts from an accurate and relevant characterization of the assumed paradigm target: typical adult humans participating in ordinary inter-personal relationships.

3.3 A Modest Empirically Informed Account of Moral Agency

How should we specify the requirements of moral agency, given a practice-focused approach? It goes without saying that how one specifies the details of these requirements matters greatly for whether, and to what extent, moral agency can be extended to nonstandard cases, such as nonhuman animals and artificial intelligence entities. This, in turn, may have significant practical and normative implications. However, it is important to remember that how one answers these questions matters in the first instance for whether one's chosen concept of moral agency accurately represents the features (behaviors, mental states, and processes) of actual human beings. This, in turn, of course, has implications for the relevance

[31] See McKenna (2013).

of the account as well as the validity of any following assessments about the possibility of moral agency in nonstandard cases.

In this section, I will follow the lead of Strawson and other practice-focused accounts to distill an important, but under-appreciated, desideratum for a theory of moral agency. Namely, that the features, and subsequent requirements of moral agency, be set to a level encompassing *typical* humans participating in *ordinary* everyday inter-relational practices. Or, put differently, that the requirements be set at the level of typical rather than exceptional performance. As I will attempt to show, this desideratum is equally important for Strawson's original reconciling project as it is for the present thesis' objective of determining whether there are any nonhuman moral agents. The latter, comparative, objective, however, makes particularly explicit the necessity of a more systematic, and empirically informed, practice-focused approach. I will therefore discuss these objectives in turn, how they each make explicit the same desideratum, and describe how resolving the risk of *anthropofabulation* (exaggerated assumption of human moral agency) (Buckner, 2013) has informed each of the four papers.

3.3.1 *The Reconciliation Project*: Internal Coherence

According to the first objective, the motivation for looking at and considering our moral responsibility practices is to arrive at an account that can reconcile determinism and moral responsibility (J. Campbell, 2017).[32] This route assumes what I call the internal coherence desideratum. According to this desideratum, the nature and conditions of moral responsibility should make sense of and be informed by the practical stance of holding people responsible. Hence, a good account of moral agency (and its requirements) should follow entirely from the logic internal to everyday attitudes and behaviors of moral responsibility.

This understanding appears to most closely follow Strawson's original suggestion: we should account for the nature and conditions of moral responsibility in terms of what it is actually like to be engaged in the attitudes and practices in question. A good explication of moral agency is thus assumed to start from an account of what we can learn from the experience (attitudes, assumptions, perceptions, beliefs) of holding responsible, as well as when one excuses or exempts.

As mentioned, the general interpretation and verdict of both Strawson and others is that our inclination for the reactive attitudes is grounded in the

[32] J. Campbell refers to this as "Strawson's Reconciliation Project" (2017, p. 32).

"human commitment to participation in ordinary inter-personal relationships" (Strawson, 1962/1982, p. 68). The truth of determinism therefore makes no difference to our general propensity for these expressive emotions. What is more, the moral emotions track features in conduct that manifest the degree of regard or concern that someone has for us and others. When we feel and express, say, resentment toward someone for having poked us in the eye, we are therefore not concerned with free will. Instead, reactive attitudes are occasioned by indications of lack of regard and express the basic expectation or demand for sufficient regard. The object of the emotional evaluations, constitutive of holding responsible, are therefore fully compatible with the truth of determinism.

This project of reconciling moral responsibility and determinism assumes everyday social interactions and attitudes as a naturalistic basis for formulating a compatibilist defense of moral responsibility. Our moral responsibility practices in general, and our reactions to perceived harms or injuries in particular, are taken to demonstrate that moral agency entails being eligible for the demand or expectation for sufficient regard or concern. Since "normal" humans are assumed to be inescapably liable to these attitudes, and since these attitudes imply conditions and requirements within the reach of "normal" humans, moral responsibility and agency are compatible with determinism. It makes no sense to question the rationality or justification of our moral responsibility practices on metaphysical grounds. Any such objections are beside the point.

Hence, the success of the reconciling project depends on the relevance and validity of the descriptive claims about our moral responsibility practices. For instance, it must be true that the features required for manifestations of quality of will do not presuppose libertarian free will. In addition, the assumed features also must accurately represent the ordinary interactions and features of *normal* or *typical* humans. That is, it must be true that (typical adult) humans are susceptible to resentment, indignation, or guilt, among other attitudes, in response to behaviors perceived as indicative of, or manifesting substandard regard. In addition, the features required for quality of will need to be widely prevalent in humans.

If the reactions and ascriptions of moral responsibility to which we are allegedly inescapably prone, assume or imply features and requirements that prove to be impossible, or very difficult to attain, or rarely met, we seem to have to reconsider the relevance of those conditions. That is, given certain requirements of moral agency, we may have reason to question the rationality and coherence of our moral responsibility practices. For example, the appropriateness of adopting the participant stance as our default perspective.

Despite the importance of an accurate and relevant characterization of the assumed features and requirements of moral agency for the success of Strawson's naturalistic argument(s), the question of how to certify such accuracy represents an under-discussed issue in most practice-focused accounts. However, *anthropofabulation* - the bias of assuming "inflationary answers to semantic questions on the basis of insufficient evidence" (Buckner, 2013, p. 863) regarding human capacities and features, has recently gained increased recognition in philosophical discussions of cross-species comparative psychology. As have suggestions for possible correctives. This brings us to the second objective of adopting a practice-focused approach.

3.3.2 The Goal of Valid Comparisons: *Hume's Dictum*

The question of accuracy and relevance when determining the features and requirements of moral agency is likewise fundamental for the possibility of making well-grounded comparative assessments. In order to consider the possibility of moral agency beyond typical adult humans in a way that is valid and practically relevant, the comparison needs to start from an accurate understanding of how such agency actually figures, and what it requires, in real life settings. In other words, one's assumed baseline needs to (at least) accurately represent and be applicable to, the alleged paradigm target.

An increasingly recognized problem for comparative assessments of psychological traits across species is *anthropofabulation*, which denotes the tendency to set the baseline of comparisons at the level of "exceptional human performance" (Buckner, 2013, p. 861). This bias stems from the common tendency of people to "confabulate about the complexity of their own performance" (Buckner, 2013, footnote 7) and to therefore "tie competence criteria … to an exaggerated sense of typical human performance." (2013, p. 853). The problem of these biases for cross-species comparisons is that they make one focus "on rarified human abilities without adequate theoretical justification" (Buckner, 2013, p. 863), which may lead one to underestimate the abilities of nonstandard cases.[33]

A corrective to anthropofabulation is suggested to be found in *Hume's Dictum* (Buckner, 2013, p. 864). This principle demands that "we set competence criteria for vaguely-defined capacities not to the highest ranks of human performance, but

[33] We will have reason to revisit this concept and its corrective in the next chapter (4).

rather only to the typical performance of children and the folk." (Buckner, 2013, p. 866). Applied to the question at hand, Hume's Dictum implies that a good account of moral responsibility and agency should be relevant and applicable to the assumed paradigm target: the ordinary practices and features of typical adult humans.

In this sense, Hume's Dictum echoes the mentioned applicability premise implicit in Strawson's naturalistic argument. Whatever features are required for moral agency, these need to be such that they are prevalent in one's assumed paradigm example. Given that the argument is intended to vindicate our ordinary everyday moral responsibility practices, then the vast majority of human adults, (and to varying degrees, perhaps also adolescents and children) would seem to need to be accommodated for.

From the over-arching research question in this thesis - the possibility of extending moral agency beyond (typical adult) humans – Hume's Dictum seems to gain importance for reasons similar to the ones making it crucial for the general Strawsonian account of moral agency. Anthropofabulation is a cognitive bias to which humans are prone and which may distort the comparison between the paradigm and the nonstandard cases of moral agency. As such, it cannot be countered by means of mere armchair recollections, thought experiments, or folk psychological intuitions. To the contrary, Hume's Dictum calls for the formulation of independent standards and "objective, empirical assessment of whether those criteria have been satisfied" (Buckner, 2013, p. 868).

Hence, the corrective principle demands an external point of view external to that of single arbitrary human participants in moral responsibility practices in order for one to describe the relevant behaviors, mental states, and processes that constitute these practices. This is needed to ground an accurate description of the behaviors and mental states and processes underpinning everyday interactions assuming moral responsibility and agency. From this characterization, one may then infer conditions and requirements of moral responsibility and agency that reflect, and apply to, the assumed paradigm target, and which therefore offer a valid basis for comparative assessment beyond that paradigm.

I will now consider how one may satisfy the applicability desideratum and how doing so requires accuracy in terms of the states, processes and behaviors involved in moral responsibility practices. But I will also suggest that considering applicability and accuracy may additionally serve to improve accuracy in terms of our individual subjective experiences of what goes on in these practices. I will also explain how my approach and findings in each of the four papers, as well as in the

following chapters, are informed by the methodological stance and desiderata discussed in this section. This section is then concluded with some reflections on a fundamental, but possibly problematic, tension between the desideratum of accuracy and the desideratum of normative guidance.

3.3.2.1 Applicability and Accuracy: Practices and Behaviors

I claim that there is a tendency in the moral responsibility and agency literature in general to limit inquiries about moral agency to questions about eligibility for moral responsibility. In addition, many accounts seem to assume that there is one unifying mechanism underlying the inclination and capacity for participation in moral exchanges. In other words, many accounts make the *a priori* assumption that there is one specific psychological mechanism or process that underlies moral agency. However, this notion runs the risk of heavily biasing data selection (Sneddon, 2005; Westra & Andrews, 2022).

If an account overlooks, downplays, or ignores, principal behavioral and cognitive aspects of participating in moral responsibility practices, this could potentially compromise any subsequent assumptions or arguments about the nature and boundaries of moral agency. Because of this, I argue that the objective of comparative validity requires us to also consider and account for the *whole range* as well as the possible *diversity* of behaviors and cognitive underpinnings involved in moral responsibility practices. Consequently, I will here discuss and question the first of these assumptions, concerning range, and will return to the assumption about monism and homogeneity in the next chapter (sec. 4.1.7).

A common way to discuss the possibility of moral agency in nonstandard cases is to ask whether, for example, small children, adults with allegedly moral agency-undermining conditions, nonhuman animals, or artificial intelligence-based machines or software, are appropriate targets or objects of moral appraisal, such as blame or praise, etc. This approach, however, assumes a limited conception of moral responsibility practices, and, consequently, of participation in such practices. Moreover, we also risk omitting subjective aspects of our practices, such as attitudes, feelings, perceptions, that may be important for understanding moral responsibility and agency.

In Paper III I therefore suggest attending to a broader range of moral responsibility reactions and responses to account for overlooked aspects or dimensions of moral agency. This proposal is in line with Strawson's original suggestion: we should consider our moral responsibility practices to learn about the nature, conditions, and requirements of moral responsibility and agency.

However, the suggestion in question follows more closely Shoemaker's suggestion to "take much more seriously than has been done before what the *whole range* of our responsibility responses consists in, what precisely these responses target, and what capacities they presuppose." (2013, p. 102; 2015 [emphasis added]).

The practice-focused approach to moral agency provides an excellent basis for extending our understanding of such agency accordingly. A careful analysis of our moral responsibility practices makes explicit that moral agency involves more than participating in the position of recipient or target of reactions and ascriptions of moral responsibility, such as resentment, gratitude, blame, or praise. Looking at some everyday scenarios involving standard as well as nonstandard agents makes explicit that agents likewise participate as sources of reactions and makers of ascriptions of moral responsibility in terms of, for example, expressing resentment. I, therefore, suggest distinguishing between a defendant and a claimant participatory position or role and argue that some parties that fail to fulfill the defendant role may nevertheless fit a claimant role.

In this way, the way we understand and conceive participation in moral responsibility practices is of great importance for questions about the boundaries of participation. Given a practice-focused conception of moral agency, a more comprehensive appreciation of participation is directly relevant to questions about moral agency in nonstandard cases. Recognizing the claimant position in a moral exchange as an essential dimension of participation in moral responsibility practices appears to have the potential to (radically) extend the scope of possible moral agents.

However, while the claimant-defendant distinction may be of theoretical interest, it also makes explicit possibly important practical and normative implications. While some of these implications are briefly touched upon in Paper III, a particular normative implication is given center stage in Paper IV. There, I argue that seeing or exempting someone as a moral claimant seems to dispose the stance-taker's sensitivity and responsiveness very differently toward the being in question. These differences, in turn, make explicit distinct other-regarding perspectives and provide normative reasons to refrain from a wholly objective stance to moral patients that fail to meet the requirements for a defendant moral agency role. I will return to and elaborate on these arguments in Chapter 5.

3.3.2.2 Applicability and Accuracy: Psychology

The second issue I want to discuss concerns the assumed relevance of certain proclaimed requirements of moral agency. As mentioned, the naturalistic

compatibilist strategy inherent to a practice-focused approach seems to already assume that a good account of moral agency is applicable to the behaviors and psychology of typical adult humans participating in ordinary inter-relational practices.

The consideration of psychological applicability figures in Papers I and II, where the relevance of some traditional requirements, such as phenomenal consciousness and metacognition, appealed to in skeptical verdicts about the possibility of moral agency in nonhuman entities, are questioned. In the first paper, Munthe and I raise doubts about the relevance of phenomenal consciousness as an intra-individualistic and vague criterion, but nevertheless often put forth in support of categorically rejecting the possibility of artificial moral agency. In the second paper, I question the relevance of conscious deliberation and moral reflection for moral agency, a requirement typically used against the possibility of animal moral agency.

Here, I will expand on how the discussion in Paper II is informed by the motivation to conduct valid assessments of moral agency in entities outside the assumed paradigm. This objective pushes to the forefront the desideratum that a good account of moral agency needs to be relevant and applicable to the features and moral responsibility practices of those already assumed to be moral agents. I will question the relevance of moral knowledge in terms of awareness of moral significance as a requirement of moral agency. I will, however, follow the view of most practice-focused accounts that moral agency in terms of manifesting quality of will requires being sensitive and responsive to moral considerations.

As discussed earlier, as well as in Paper II, a disadvantage of capacity-focused accounts is that they tend to assume over-intellectualized conditions of moral responsibility. According to many capacity-focused accounts, an agent is praise- or blameworthy to the extent that they willfully and knowingly act on right or wrong reasons (Korsgaard, 2006; Dixon, 2008). Moral agency is therefore believed to require metacognitive states and processes, enabling the agent to engage in conscious (moral) reflection. In particular, being morally responsive is taken to require moral knowledge or awareness in the sense of being explicitly aware of the moral significance of an action or choice (Haji, 1997; Zimmerman, 2002, Levy, 2011). This is often referred to as *de dicto* moral awareness and is contrasted to *de re* moral awareness, namely, awareness of the right- or wrong-making features of a situation. This latter type of awareness does not, then, require that the agent believes or is aware of the moral significance of these features (Rudy-Hiller, 2022a).

For instance, Dixon claims that "[w]hat is central to emotional motives like compassion is not merely that the agent is moved to perform morally right actions, but that she understands that the action is virtuous and performs that action for the sake of virtue and not for some other reason" (Dixon, 2008, p. 76). Moreover, "[a]t the minimum, to be morally appraisable one needs to understand something about the concepts of morality in addition to being emotional and cognitively responsive to moral particulars. Since even trained animals are unable to do this, they are exempt from this sort of minimal responsibility ascription" (Dixon, 2008, p. 199).

However, this and similar views are not based on an accurate understanding of *our* moral responsibility practices and the agency implied therein. Nor are they empirically informed as to what actual psychological states and processes real humans seem to make use of when navigating moral considerations and participating in moral responsibility practices. As a result, any conclusion about the possibility to attribute moral agency to nonhuman entities based on an incorrect understanding, will be compromised.

Let us start by looking at our practices of reacting and ascribing moral responsibility and what they can tell us about the reasons in favor of a condition of de dicto moral awareness. According to the moral awareness condition mentioned above, an agent is blameworthy to the extent that they are aware (either at the time of the action, or dispositionally) that the action is right/wrong. That is, for the eye-poker to be blameworthy, they need to have known that poking people in the eyes is wrong (or, for instance, that causing people discomfort is wrong) to be a fitting or appropriate target of blame.

However, when we blame someone, we do not appear to assume that the agent in question needs to be aware of the moral significance of their action. If the eye-poker had the requisite factual awareness, specifically, if they knew that they were poking someone's eye and knew that poking someone in the eye causes discomfort, we seem to think that they are blameworthy. Hence, looking at our practices and the attitudes and behaviors involved, we do not seem to treat moral ignorance as a general excuse for moral responsibility. Considering the mentioned requirements as representative of human participation in moral responsibility practices hence runs the risk of committing to anthropofabulation. Assuming a quality of will condition not including this feature thus seems to have a better prospect of avoiding this flaw and to make better sense of our everyday ascriptions.

To avoid anthropofabulation about moral agency, and be able to make valid comparative assessments, we need to inform our understanding, and subsequent

requirements, on, not only armchair recollections of the nature of participation in moral exchanges, but also on empirical data. An account that sets out to do just this, is Nomy Arpaly's quality of will theory, according to which "deliberation is given far more prominence in moral psychology than its position in daily life would suggest" (Arpaly, 2002, p. 21). According to Arpaly, we do not seem to have any good reason to think that the type of reason or rationality required for moral agency implies that one consciously thinks about, or reflects on, what to do.

Similarly, responsiveness to reasons does not appear to require conscious reflection or awareness of one's beliefs, desires, or reasons. Contrary to folk psychological assumptions, much of human behavior, including decision-making, seems to be controlled by nondeliberative, unconscious, and simple cognitive processes (Sie & Wouters, 2010). Even "apparently complex behavior" can many times be explained by reference to "elementary mechanisms" (Shettleworth, 2010, p. 480). And the role of feelings and emotions in acting rationally appears to be much more important than what many have thought to be the case (Damasio, 1994).

Noncognitive states can act as subtle cues, giving us access to background knowledge by functioning as "markers", pointing us in different practical directions. Clinical cases of patients who lack "somatic markers" (Damasio et al., 1991) due to brain injury seem to show that practical rationality is greatly undermined without them. These patients are completely unimpaired with regard to intellect, memory, knowledge base and general problem-solving abilities. But they are seriously impaired regarding personal decision-making (Bechara et al., 1994).

If rationality requires that an agent is caused to act by her conscious deliberation, or if acting for reasons requires such deliberation, "we would have to call people rational considerably less often than we do" and "we would find that it is uncomfortably rare for people to act for reasons" (Arpaly, 2002, p. 51). Similarly, if moral agency requires this kind of deliberation, we will have to call people morally responsible much less often than we do. Hence, Arpaly defends an account of reason responsiveness in general, and moral responsiveness in particular, that does not require deliberate or conscious deliberation. An agent can respond to reasons without knowing, or being aware of, that she is. That is, an agent does not need to entertain any belief (whether conscious or unconscious) about a reason to still be moved by that reason. Likewise, an agent can be a moral agent, without knowing, or being aware of, the moral reasons that move her.

Hence, while cases involving conscious moral reflection and reasoning may appear intuitively significant and commonplace, we seem to have good reason to

be skeptical of such intuitions. Humans have been shown to routinely exaggerate "our own intelligence, rationality, and reflective prowess" (Buckner, 2013, p. 860). Psychological studies demonstrate that we are frequently overconfident in our psychological abilities and that we ignore or distort counterevidence. What is more, data indicates that we (sincerely) provide reasons and justifications for our actions, despite those actions having been due to "whims, heuristics, or situational factors" (Buckner, 2013, p. 860).[34]

In this way, empirical data seems to undermine the validity of conceiving moral responsiveness in terms of conscious reflective endorsement. Such *de dicto* awareness just does not seem to be very prevalent in, or relevant to, the way that actual human beings navigate the moral landscape. Hence, if sensitivity and responsiveness to moral considerations requires entertaining beliefs about the wrongness/rightness of an action, even typical adult humans would very seldomly be eligible for ascriptions of moral responsibility (Arpaly, 2002; Markovits, 2010). This undermines the credibility of a number of moral agency requirements based on an intuitively assumed condition of *de dicto* awareness of moral significance, such as, for example, metacognitive states and processes or conceptual knowledge. Consequently, any conclusions drawn regarding the (im)possibility of moral agency in nonstandard cases are compromised when these requirements are called into question.

In contrast, a benefit of practice-focused views is that they generally do not treat *de dicto* moral awareness as a prerequisite for moral responsibility. Our practices of reacting to perceived right- or wrongdoing and good and bad conduct with praise or blame are instead assumed to track the agent's quality of will rather than their awareness of moral significance per se. That is, practice-focused accounts assume quality of will, and the ability to notice and react to such will, as the basic condition of moral responsibility. An action expresses good will if it arises from sufficient or proper concern and ill will if it stems from insufficient or lack of concern.

In this way, according to a quality of will condition of moral responsibility, an agent does not need to believe that their action is right or wrong for their action to express their quality of will, and thus for them to be blame- or praiseworthy (Fields, 1994; Arpaly, 2002, 2015; Harman 2011; Talbert, 2013, 2017; Mason 2015; Björnsson 2017; Weatherson, 2019). Hence, understanding acting with good will

[34] See also Ariely (2009, 2012), Bermúdez (2003), Gilovich and colleagues (2002), Malle and colleagues (2007), and Malle (2011).

or lack of concern as being (un)responsive to moral considerations *de dicto*, appears to unnecessarily over-intellectualize moral agency.

This is not to say that deliberate moral reasoning never occurs or that it lacks significance altogether. There are certainly times when typical adult humans engage in explicit and linguistic moral exchanges. Likewise, there are situations in which we find ourselves engaging in deliberative moral inquiry. And there are certainly occasions in which conscious reflection facilitates, rather than undermines or is superfluous to, moral evaluations. However, given that the lion's share of the mental states and behaviors involved in participation in moral practices does not involve *de dicto* awareness of moral properties there seems to be little in favor of treating deliberate, explicit, and linguistically mediated, moral reasoning as a constituting paradigm feature, or the basic threshold, of moral agency.

3.3.2.3 Applicability and Accuracy: A Remaining Tension

Lastly, I want to say some words about a possible remaining issue with the accuracy desideratum. Accurately determining how, and why, we include or exempt others, seems to suggest that ascriptions of moral responsibility and agency do not always track conditions or requirements often assumed to be relevant for moral responsibility or agency. For instance, descriptive assessments seem to show that even paradigm moral agents rely much less than commonly assumed on explicit moral reasoning.

These incongruencies between theory and folk psychological assumptions on the one hand, and actual practices and the result of systematic empirical inquiry, on the other, speak in favor of an empirically informed, modestly construed, practice-focused account. However, they also make explicit a tension within any practice-focused approach. The descriptive aim of accuracy seems to stand in tension with the prescriptive objective of normative guidance.

If observations show that people often, say, include or exempt other agents for reason x, and reason x fails to match theoretical or pre-theoretical assumptions of what a valid such reason is, accuracy seems to pull in a different direction than fittingness or appropriateness. In other words, the task of accurately representing the features and practices of real human beings does not necessarily lend itself very well to the aim of deriving normatively justified standards of moral responsibility and agency.

How can one tackle this seeming "dual burden" of a practice-focused theory of moral agency (Argetsinger & Vargas, 2022, p. 31)? I return to this question briefly in the next chapter and sketch a possible pathway to a solution in Chapter

6. I believe that the descriptive-prescriptive desiderata can be reconciled by following a particular account of the content and function of blame and similar reactive attitudes. This suggestion draws from the communicative emotion account of blame, discussed in Paper III. This account will then be paired with, and bolstered by, instrumentalist considerations.

3.4 Concluding Remarks: Moral Agency as a Social-Normative Competence

Given a practice-focused approach, moral agency just is and requires what is needed for being a participant in moral responsibility practices. This means that any requirement needs to reflect and be relevant to the nature of these practices and our participation in them. I have therefore argued that a good account of moral agency needs to accommodate the variable features, skills and practices of real participants in paradigm moral responsibility practices.

Everyday social life is immersed in simple, non-deliberative, and many times non-linguistic, moral exchanges. The cafeteria and Food Stand examples described in Paper II and III, respectively, are intended to illustrate such commonplace examples involving typical adult humans. The examples involving canid social play interactions in Paper II, and the suggested examples of moral address from a toddler and a dog in Paper III are intended to show analogous behaviors in non-standard cases.

The reason for preferring a practice-focused approach over a capacity-focused approach is, therefore, that the former lends itself better to an empirically informed and theoretically modest account of moral agency. It is, in other words, better suited to align with current data about the psychology and behavior of actual humans, to account for everyday interactions where moral agency and responsibility are assumed, and to provide a valid framework for considering the possibility of moral agency in nonstandard cases.

A conception of moral agency that satisfies the mentioned desiderata, asserts that, in practice, such agency involves qualifying as a participant in moral responsibility practices. This involves being *liable to* the reactive attitudes and being *eligible for* the deployment of such attitudes. Such liability and eligibility imply sensitivity and responsiveness to moral considerations, where this may be construed in terms of, for instance, caring for or respecting the welfare, interests, rights, wants, et cetera of others, but also of oneself. Some such considerations are connected to standards, norms, or expectations for right, good, or appropriate

conduct. Given the centrality of social maintenance behaviors, participation in terms of liability and eligibility likewise involves an affective-communicative competence. Hence moral agency should be conceived in terms of a set of competencies to recognize and respond to morally significant features of situations and to engage others, and be engaged with, in moral exchanges.

4 Nonstandard Moral Agency

While various accounts of moral agency differ in their stated requirements, what is typically referred to as the standard template for moral agency is a typical adult human being. In other words, a human being beyond childhood and adolescence, whose mind functions in ways considered expected or normal. For example, Wolf writes that "human adults of normal mental health and intelligence" are in the "first class" of agents "to whom reactive attitudes are appropriate" (2015, p. 131), and Rosen states that "we know" that "normal competent adults" are "on the hook" with regard to moral blame (2015, p. 74).

Although the mentioned standard is not always this explicitly stated, it is easily inferred by how the notion of moral agency is introduced by pointing out examples of typical agents or beings who are assumed to fall outside or stand on the margins of the domain of moral agents. For instance, Wolf also writes that "we adult human beings can be responsible for our actions in a way that dumb animals, infants, and machines cannot" (1987/2013, p. 332). And Hakli and Mäkelä (2019) believe that "robots as programmed artifacts cannot be moral agents responsible for their actions" (2019, p. 261). In this way, small children and nonhuman animals, and sometimes autonomous and advanced machines, are examples of entities and agents assumed to be exempt from ascriptions of moral responsibility, and thus to fall outside the realm of *moral* agency.

In addition to small children and the mentioned nonhuman cases, there are also many groups of adult humans assumed to have diminished moral agency due to certain putatively moral agency-undermining traits, conditions, or disabilities. These are, for example, people suffering from late-stage dementia, impaired empathy, or obsessive-compulsive disorder (Haksar, 1998; Talbert, 2022). However, *marginal* (Shoemaker, 2015), disputed, or typically exempted, cases, like those mentioned, are also a topic for debate in the philosophy of moral agency.

As previously mentioned (2.3, 3.2.4, 3.3), a prominent assumption among adherents of Strawson as well as others is that our inclination to either include, exempt, or feel ambivalent toward agents as eligible for moral reactions tracks or reflects the conditions and boundaries of moral responsibility and agency (Strawson, 1962/1982; Watson, 1987/2004; Wallace, 1994; McKenna, 2012;

Shoemaker, 2013, 2015; Rosen, 2015). While excuses may tell us something about the conditions of moral responsibility for particular acts or omissions, exemptions are assumed to reveal something about the conditions of moral agency in general. If there are certain agents that regularly fail to evoke, or toward whom we regularly modify or inhibit, our moral responsibility reactions, this may indicate that these beings lack something of importance for moral agency. A popular strategy, not least within a practice-focused approach, has therefore been to discuss and compare standard and nonstandard cases in the hopes of gaining insight into the features, properties, or powers that ground moral agency.

The aim of this chapter is to account for and analyze discussions about nonstandard cases in the moral agency and responsibility literature. I do this by spelling out and evaluating various suggestions about why we exempt, or feel ambiguous to, nonstandard cases. The first section provides an overview of some common examples of nonstandard human cases in the moral agency and responsibility literature, along with the potential deficiencies or requirements suggested to be made visible by our attitudes and behavioral inclinations in each case. The second section continues with an overview of discussions concerning the two cases of nonhuman nonstandard cases at the focus of this thesis: artificial intelligence entities and nonhuman animals. These debates will be compared and argued to highlight partially distinct questions and themes.

The last section outlines first some general reasons commonly provided for taking the objective stance and then some initial worries or issues about exemptions. I conclude this section by turning to the arguments developed in Paper III, namely that philosophical accounts considering reasons for or against exemptions tend to focus on the defendant dimension of participation. As such, they overlook or downplay an additional, and fundamental, way in which one can be included in or exempted from moral responsibility practices: namely, as a moral claimant. By considering this aspect, I argue, we may find that we need to redraw the boundaries of moral agency. What is more, the claimant-defendant distinction serves as a basis for the next section, where I highlight possible links between moral patiency and moral agency.

4.1 Nonstandard Cases of Human Moral Agency

This section sets out to account for some popular nonstandard human cases by discussing the responses assumed to be appropriate or fitting toward such agents,

and the suggested implications of these responses. I conclude this section with a critical discussion of these verdicts in light of the previously suggested desiderata (3.3).

4.1.1 *The Psychopath*

A nonstandard human case at the focus of much philosophical discussion is *the psychopath* (see Shoemaker, 2010, 2015; Kennett, 2002). The psychopath is typically assumed to be "characterized by extreme egocentricity and impulsivity, by a pronounced lack of remorse and empathy, and by a persistent tendency to disregard the effects of one's actions on others" (Talbert, 2008, p. 518).[35]

The reason for the recurring presence of this particular example in the literature, is argued to be due to the *ambivalent* (Watson, 1987/2004) or contradictory intuitions of moral responsibility and blameworthiness that the notion of this type of agent allegedly produces. While many other nonstandard cases, such as small children and nonhuman animals, are typically taken to lack the intellectual capacities of typical adult humans, "[w]hat is especially puzzling and problematic about the case of the psychopath is precisely that these individuals appear "normal" and "mature" in respect of rational self-control" (Russel, 2004, p. 297) as well as "mental state attribution" (McGeer, 2008, p. 230). Hence, the psychopath has been assumed to be a case that may reveal a good deal about the nature and requirements of moral agency.

The features assumed to indicate that the psychopath is morally incapacitated are commonly thought to be their "lack of receptivity to moral reasons" and their "lack of understanding that someone else's interests provide noninstrumental reasons for acting" (Nelkin, 2015, p. 361; see also Talbert, 2008; Watson, 2011, 2013; Kennett, 2002; Shoemaker, 2015; Levy, 2007; Mason, 2017). As such, the psychopath seems to make explicit the (in)sufficiency of rational capacities, such as practical reason and mindreading, for moral agency (Russel, 2004).

---

[35] Although authors in these debates regularly refer to what they call *psychopaths*, this particular terminology does not figure in any of the two established systems for classifying mental disorders: the World Health Organization's International Classification of Diseases (ICD) and the American Psychiatric Association's Diagnostic and Statistical Manual (DSM). In addition, it is not clear to what extent the examples in these discussions accurately track diagnostic criteria or reflect features of the diagnoses probably closest to what is referred to as psychopathy: *dissocial personality disorder* (ICD) and *antisocial personality disorder* (DSM). For the purposes of this chapter, however, I will assume the terminology employed in these discussions to account for and analyze views on the moral agency of human agents displaying the mentioned corresponding features.

A common type of claim is therefore that the psychopath lacks, or is impaired with regard to, a specifically *moral* competence (see 3.2.2 and 3.2.4) necessary for grasping and responding to moral considerations. For instance, Nichols (2001) claims that what psychopaths lack is a "Concern Mechanism" (2001, p. 426). This mechanism responds to cues indicating distress in others by producing an affective response that motivates altruistic behavior. Lacking this mechanism, Nichols believes, explains why psychopaths fail to distinguish between conventional and moral violations. Since the psychopath is defined to lack this competence, they can at most be motivated to abide by moral principles or considerations for instrumental reasons, such as the desire to avoid sanctions (similar to Russel's (2004) example about Jill who is incapable of true fear). A different argument is made by Levy (2007) who suggests that the psychopath lacks moral understanding and, in turn, is incapable of the type of *control* necessary for moral responsibility.

A more radical position is found in the claim that the psychopath makes explicit a link between "prudential and interpersonal concern" (Watson, 2013, p. 287). According to this suggestion, what the psychopath lacks is not just concern for others, but the more fundamental capacity to "form any extended and coherent conception of his own or others' ends, and therefore of the ways in which those ends generate and sustain reasons over time" (Kennett, 2002, p. 355). This incapacity to "value certain ends" (McGeer, 2008, p. 247) in turn renders the psychopath unable to invest in, and thus normatively commit to, any ends, their own or others (Kennett, 2002; McGeer, 2008; Shoemaker, 2015, Ch. 5). As such, the psychopath is, in a sense, argued to be evaluatively indifferent in general, over and above immediate whims or impulses (McGeer, 2008, p. 254; Kennett, 2002, Watson, 2013; Shoemaker, 2015, Ch. 5).[36]

The mentioned deficiencies, many argue, serve to exempt the psychopath from moral responsibility. It would be unreasonable and maybe even unfair to subject them to the demands and harms of blame or to otherwise hold them responsible (Russel, 2004; Nelkin, 2015; McGeer, 2008; Fine & Kennett, 2004; Kennett, 2002; Murphy, 1972; Wolf, 1987/2003; Shoemaker, 2010; Levy, 2007). Others, however, argue that while psychopaths indeed are morally deficient, they are not off the hook regarding moral responsibility. For instance, some claim that the psychopath is capable of guiding their conduct by reasons in general, or that their conduct still manifests their quality of will (Scanlon, 1998; Talbert, 2008; Greenspan, 2003; A.

[36] Note, however, that this deficiency is sometimes put in terms of "an affective deficit" (Kennett, 2006, p. 70; see also Nichols, 2002; McGeer, 2008) and other times as a "rational shortcoming" (Kennett, 2006, p. 70).

M. Smith, 2008). As such, the psychopath is argued to be open to blaming responses.

Some suggestions regarding the moral agency of psychopaths are located somewhere in the middle between the mentioned positions. Some proponents of pluralist accounts of moral responsibility and agency argue that, since psychopaths are capable of "intentionally or willingly" harming others (Watson, 2011, p. 317), they are eligible for moral assessments about character (attributability). However, since psychopaths lack the capacity for perceiving the normative perspective of others (their cares, commitments, emotions) as providing putative reasons (Shoemaker, 2015), they are not appropriate targets of the demand for regard (accountability) (Watson, 1987/2004, 2011; Shoemaker, 2015, 2013, 2011).[37]

While I am inclined to concur that the psychopath indeed seems to be missing something vital to moral agency, the normative and practical implications of this are not as clear. Assuming that blame can play various functional roles, it may be reasonable to blame a psychopath for harming us despite their assumed inability to truly recognize or respond appropriately to blame. For instance, following a broad understanding of blame as a form of protest or a way of standing up for oneself. holding psychopaths responsible can be reasonable since their harmful "actions can express offensive judgments that we are interested in rejecting and standing up against" (Talbert, 2012, p. 106).[38]

4.1.2 Agents with Unfortunate Upbringings

A related but distinct kind of nonstandard human case concerns agents who are "peculiarly unfortunate in formative circumstances" (Strawson, 1962/1982, p. 79): in other words, agents who are morally deprived due to abuse, neglect or misguided upbringings.

A recurring example in the responsibility literature concerns real-life kidnapper and murderer Robert Harris, first mentioned in the responsibility debate by Watson (1987/2004). Watson argues that learning about Harris' horrible upbringing and childhood trauma appears to mollify our reactive attitudes toward him. However, exempting someone from moral responsibility by appeal to some "deviant causal history" (Shoemaker, 2015, p. 192) or "moral luck" (Russel, 2011, p. 219), Watson argues, moves us beyond the resources of Strawsonian compatibilism.

---

37 Nelkin (2015) explicitly denies the attribution of moral vices or virtues to psychopaths.

38 See also Talbert (2021).

What is more, if resentment and indignation express the basic demand for good will in our moral responsibility practices, and the participant stance presupposes the possibility of such moral address, "then the paradox results that extreme evil disqualifies one for blame" (Watson 1987/2004, p. 235). While the conduct of the extremely evil seems to present paradigm examples of blameworthy conduct (like in the case of cold-blooded murder), these agents appear to lack the capacity for being motivated by moral considerations and thus are not someone to whom moral address seems appropriate. In this sense, agents of extreme evil appear to be exempted from moral agency by definition, argues Watson.

Wolf (1987/2013) provides another take on the issues presented by unfortunate formative circumstances. She argues that cases involving deprived childhoods and agents who live in misguided societies, highlight the need for supplementing Real Self views of moral agency (see also previous chapters: 2.2.1, 2.2.3, and 3.2.2) with a moral competence condition.[39] Wolf illustrates her point by using the fictional example of JoJo, a boy who is brought up by a sadistic dictator and comes to adopt the vile values and behaviors of his father. Applying the conditions of Real Self views, Wolf argues, would lead us to the implausible conclusion that JoJo's actions express his deep beliefs and values (or his quality of will, see Shoemaker, 2015, Ch. 7) and that he therefore is a moral agent.

However, Wolf argues, learning about JoJo's upbringing and the misguided society he lives in seems to make us inclined to exempt him from blame. How can we make sense of these intuitions, on one hand, and maintain the appropriateness of a general participant stance, on the other? Wolf's suggestion is to add to moral agency a condition of "sanity", that is, "the *ability* cognitively and normatively to understand and appreciate the world for what it is" (1987/2013, p. 338). Hence, because JoJo's beliefs and values are "unavoidably mistaken" (1987/2013, p. 336), he lacks the moral competence required for being fully responsible for them, Wolf argues.[40]

At the same time, several authors reject unfortunate formative circumstances as an example of a wholly exempting condition. The formative circumstances may be regrettable in how they have shaped, and perhaps still enforce, the preferences and values of these people. Moreover, their parents and or society may very well

---

[39] Note that Wolf uses the different term *deep-self view* (1987/2013) here to refer to the type of account that she later refers to as the Real Self View (Wolf, 1990).

[40] See also Sie (2005), who argues that Wolf and Wallace seem to be "committed to the existence of a fundamental distinction between wrongdoers and blameworthy agents, a distinction that separates the two independently of the decisive identification (or lack thereof) of the agent." (2005, p. 64).

be to blame for creating their bad characters. However, since the conduct of these agents does in fact manifest their substandard regard or vice, they are morally responsible (Talbert, 2012; A. M. Smith, 2008; Scanlon, 1998; Arpaly, 2002, Ch. 5; Moody-Adams, 1994).

Our ambivalent or conflicting intuitions regarding cases like Harris' may be due to a failure to distinguish between various forms or stages of moral responsibility. This may be because we commonly seem to conflate assessments of character with reactions to (insufficient) regard (Shoemaker, 2015), for example, or because we tend to conflate the question of whether someone is an apt object or recipient of, say, blame, on one hand, and what exactly would be an appropriate or justified response to their wrongdoing, on the other (Arpaly, 2002; Levy, 2003).[41] Or, it may be because the nature of human moral psychology makes it difficult for us to *typecast* an agent as both perpetrator and victim at the same time (Gray et al., 2012).

More nuanced, and I believe more attractive, views regarding the moral agency of people with unfortunate formative upbringings can be obtained by attending to, and teasing apart, various possible aims and values of ascribing moral responsibility. For instance, according to explicitly educational, scaffolding, or reformative approaches to moral responsibility, blame may be fitting even in cases concerning disorders affecting moral agency, since such agents may still be capable of rehabilitation and reform, and thus of becoming better or more capable (moral) agents (Pickard, 2014).[42]

Likewise, it may be the case that many people fail to be appropriately responsive to moral considerations *because* they live in societies or cultures that foster oppressive behavior. Still, blame can make sense in terms of, for instance, effecting social change (Calhoun, 1989)[43] or as a way of reminding, educating, or in other ways scaffolding the sensitivity and responsiveness of moral agents.[44]As

---

41 Interestingly, while Arpaly (2002) and Levy (2003) emphasize the distinction between blameworthiness and various types of punitive measures, they do so to explain and support opposing positions. Arpaly believes that an agent such as JoJo can be blameworthy but not necessarily open to punishment, while Levy argues that JoJo is excused from blame but not legal consequences.

42 Pickard (2014) argues that since the very nature of personality disorders makes the agent vulnerable to "feelings of rejection, anger, shame, hopelessness, and desperation" (2014, p. 8), blaming responses need to be void of anger in order to make possible their beneficial or scaffolding effects.

43 Calhoun (1989) argues that it makes sense to hold people responsible for behavior that is unwittingly sexually or racially oppressive, albeit in a detached way, to effect social change.

44 These suggestions assume ideas similar to those of modern instrumentalist accounts (see also Chapter 6). A significant difference, however, is that the former, but not the latter,

mentioned in conjunction with the case of the psychopath, blame may also have aims and values beyond contrition and reform, such as, say, affirming the blamer's self-worth (Talbert, 2012).

As we will see, arguments that appeal to finer distinctions, such as types, stages and aims, and values of reactions and ascriptions of moral responsibility, frequently figure also in discussions about nonhuman moral agency (see 4.2 below). Following this, Paper III sets out to make explicit an overlooked dimension of moral participation, and Paper IV discusses normative reasons for and against including or exempting someone as moral agent.

4.1.3 Autistic Persons

A third example of a nonstandard human case concerns the moral agency and responsibility of autistic persons (Kennett, 2002; McGeer, 2008, 2009; Shoemaker, 2015; Stout, 2016). A central theme here is to investigate the role of social cognition, mindreading, or (various forms of) empathy for moral agency, on the assumption that autistic persons differ from neurotypicals regarding these capacities or features. For example, a common view is that autistic persons rely (more heavily than neurotypicals) on explicit forms of inference when interpreting the mental states of others (McGeer, 2008, 2009; Stout, 2016; Kennett, 2002). Some claim that because autistic persons are impaired with regard to certain forms of social cognition, mindreading, empathy or self-directed reactive attitudes (such as guilt), they are not (full) moral agents (see Shoemaker, 2015; Stout, 2016).

However, many others (myself included) deny these conclusions and argue that, while autistic persons may differ from neurotypicals in some respect regarding mentalization, the significance of this for moral agency is far from clear or established. For example, some propose pluralism with regard to the features requires for moral agency and claim that empirical evidence as well as lack of consensus on what empathy, prosocial behavior, or moral agency even requires, forces us to question the view of moral agency "as an "all or nothing" set of capacities", that excludes autistic persons (Krahn & Fenton, 2009, p. 158; see also McGeer, 2008). I am sympathetic to the notion of moral agential heterogeneity and will return to this matter shortly (4.1.7).

---

differentiates full-blown paradigmatic blame from the suggested detached, rehabilitative, or educational responsibility responses.

In addition, some authors argue that because learning and following moral rules or principles does not (always) require complex or explicit mindreading, excluding autistic persons from moral agency would be a mistake (Kennett, 2002, 2006; McGeer, 2008).[45] A related, but sentimentalist, argument against exempting autistic persons from moral agency can be derived from positions that suggest the primacy of sympathy rather than empathy for moral motivation and behavior. Given that autistic persons have the affective system required for concern (Nichols, 2001), and assuming that concern and not empathy is central to moral motivation and conduct, autistic persons are moral agents. As we will see, similar sentimentalist positions about altruistic/moral motivation are popular among proponents in the animal moral agency debate.

4.1.4 Other Nonstandard Cases from the Psychiatric Domain

There are many more putative examples of nonstandard human agents from the psychiatric domain discussed in the moral agency and responsibility literature. These are, for example, people with ADHD, bipolar disorder, people who suffer from addiction, kleptomania, pyromania, obsessive compulsive disorder (OCD), Tourette's syndrome, emotional instability syndrome, eating disorders, dementia (such as Alzheimer's disease), and psychosis (Arpaly, 2002; Pickard, 2011, 2015; Pickard & Ward, 2013; Kennett, 2009; Shoemaker, 2015; Jennings, 2009; Brandenburg, 2018; Radoilska, 2023). While an in-depth discussion of every mentioned case is beyond the scope of this thesis, it might be helpful to consider and compare some suggestions in the moral agency and responsibility literature that are purportedly all-encompassing or exhaustive.

According to an autonomy-emphasizing account, what unifies many of these examples is suggested to be that the illness or disorder undermines, disrupts, or prevents the person's "unity of agency" (Kennett, 2009, p. 93). When an agent suffers from motivational deficits, compulsion, delusion, or memory loss, their "projects, large and small" are "at constant risk of derailment…" which means that they "…lack authorial control over" their life (Kennett, 2009, p. 96), and thus cannot be appropriately held responsible.

However, unsurprisingly, I am inclined toward a quality of will analysis. The responsibility-undermining features of, for example, ADHD, kleptomania, and

[45] See Batson (1994) and Batson and colleagues (2011), who discuss *principlism* as a possible alternative basis for helping behavior in humans aside from altruistic motivation.

depression are not primarily due to constitutive or correlated deficiencies in conscious deliberation, autonomy, or self-control. Rather, deviations in psychological make-up are relevant to moral responsibility to the extent that they affect the agent's responsiveness to moral consideration. The question, therefore, is whether, and to what extent, a feature explains (away) ill will or lack of concern (Arpaly, 2002, Ch. 5), such as when a feature impedes or in some way imposes significant costs on, the agent's ability and inclination to recognize and respond to moral considerations.

For instance, consider an agent, A, who suffers from social anxiety. Due to her condition, A finds it very difficult to maintain eye contact, avoids answering phone calls or replying to messages, and regularly cancels social plans at the last minute. While the mentioned behaviors might be indicative of disrespect or insufficient regard in the case of many other people, this is clearly not the case here. The severe distress that A experiences due to her social anxiety explains her tendency to avoid attention and social interaction without implicating the quality of her will.

In addition, following the practice-focused account proposed in the previous chapter (3.4), I believe that an agent can be exempted if her condition or incapacity undermines her ability to engage with others in moral exchanges. Consider, for instance, an agent, B, who suffers from paranoid personality disorder and who, due to her general mistrust of people, regularly misinterprets or misrepresents feedback or criticism. Her condition disposes her to perceive most instances of criticism in general as pre-meditated attempts aimed to hurt her. Hence, when she is blamed for, say, having failed to keep a promise, she is not capable of uptake like most other people. It is not the case that she dismisses blame just to avoid the aversiveness of guilt. Nor is it merely the case that she questions criticism or feedback on reasons of proportionality. Rather, her paranoid disposition makes it very difficult for her to perceive and seriously consider the moral message of blame in the first place.

4.1.5 Persons with Intellectual Disabilities

Another subset of nonstandard, marginal, or typically exempted, human agents discussed in relation to moral agency and responsibility are persons with intellectual disabilities. Shoemaker (2015) argues that although this group is commonly grouped together with children in the literature, this is a mistake. He claims that the majority of people with mild intellectual disability (or MID) are

generally more mature than small children in respects relevant for ascriptions of moral responsibility in the accountability and attributability senses.

However, Shoemaker argues that because people with MID "may have trouble accessing or appreciating abstract principles about mutual recognition and accountability amongst *all* members of the accountability community" or "trouble seeing (without serious prompting) how the practices of mutual accountability with which they are familiar ought to be applied to unfamiliar agents", they may not be full members in the accountability community with regard to people outside their circle of family, friends and caregivers (2015, p. 187). Shoemaker also argues that because many adults with MID are not "able to engage in abstract thought or to apply principles or information from one situation to another very well" (2015, p. 183), they have impaired answerability.

While I am inclined to disagree with the broad-brush approaches and generic verdicts like the one above, I am sympathetic to the idea that moral agency may depend on, and vary across, social contexts. In Paper III, I argue that various aspects or dimensions of moral agency may depend on, and vary with, the particular social context and interaction in which a particular agent is situated. For instance, small children or dogs may qualify as co-participants with typical adult humans in the sense of being eligible for moral responses, while failing to qualify in the sense of being eligible for moral appraisal.

What is more, I believe that typical adult humans likewise regularly struggle with recognizing unfamiliar, distant, vulnerable, different, disenfranchised, or nonhuman agents as co-participants in moral responsibility practices. This is, in fact, a central claim developed and defended in Paper IV, and elaborated on in Chapter 5.

4.1.6 Children and Adolescents

Interestingly, the use of small children as examples of nonstandard cases in philosophical discussions of moral agency is more nuanced than many of the examples assuming psychiatric conditions. For instance, even though young children are typically assumed to lack the control and knowledge[46] required for

[46] In addition, children and childhood is commonly associated with "innocence", and "identities of purity, the absence of vice, vulnerability, and a lack of knowledge or understanding of the world" (Burroughs, 2020, p. 87; see also Coveney, 1982). These cultural associations, some argue, play an important role for the ways we understand, imagine and practically assess children as moral agents (Burroughs, 2020). An interesting contrast here, however, is so-called

moral agency,[47] many authors explicitly recognize that our everyday interactions with children seem to tell a somewhat different story about our reactive attitudes.

According to some, the fact that adult humans regularly react to the conduct of young children with disapproval and annoyance can be straightforwardly explained in purely educational terms. What may look like blame in these instances, is nothing more than the application of a feigned, educational, or *as if*-stance aimed at preparing children for their future lives as full-blown moral agents (see Strawson, 1962/1982; Vargas, 2013; Sneddon, 2005). Full-blown blame would, on the other hand, be unreasonable or unfair (see, for example, Wallace, 1994).

Others, however, suggest that children can be actual participants in (some) moral responsibility practices precisely in virtue of the regulative function of blame. If children can be influenced to internalize moral considerations via (some) moral responsibility responses, they are eligible in the requisite sense (Burroughs, 2020; Brandenburg, 2019). Hence, this view emphasizes, and reflects, what I have called the communicative, or *affective-communicative*, skills of moral agency (3.4).

The position, perhaps, closest to the suggested account in 3.4, is found in views that conceive of moral agency in terms of a social-normative competence. Strawson described childhood as a "borderline, penumbral area," and where we find "progressive emergence" of moral agency (1962/1982, p. 75). Since children are psychologically less developed as well as less experienced than adults, they are generally not as morally competent. In other words, children will find it more difficult to recognize and respond to moral considerations than typical human adults. However, just like adults, children are morally responsible relative to their moral competence, and therefore should not be assumed to be wholly exempt from moral agency (Tiboris, 2014).

---

adultification of some children, that is, when some "children are perceived as being less innocent and less vulnerable, and subsequently not afforded the same protection" as other children (Davis & Marsh, 2020, p. 256). For example, racial biases lead people to meet Black children with "suspicion, assumed deviance and culpability" (Davis, 2022, p. 5; see Goff et al., 2014).

47 Very young children, such as infants and toddlers, are often thought of as clear-cut examples of responsibility exempted humans (Shoemaker, 2015). This is because young children are assumed to lack reason-responsiveness, deliberation, self-constitution, moral competence and/or self-control (Tiboris, 2014). Instead, they are thought to be moved by their appetites, impulses, and emotions (Burroughs, 2020). On one interpretation, this makes them incapable of avoiding wrongdoing and therefore not fair targets of ascriptions of moral responsibility, such as blame (see, for example, Wallace, 1994). On another interpretation, lacking these basic moral and rational competences, the actions of children, however annoying or harmful, fail to have the requisite sort of moral significance required for ascriptions and reactions of moral responsibility to be appropriate (Watson 1987/2004; Shoemaker, 2011; McKenna, 2012).

Considerations of moral agency and responsibility in older children and adolescents, make explicit the transitional nature of acquiring (full) moral agency. In this way, childhood and adolescence are taken to challenge views that assume sharp distinctions between those who are and those who are not moral agents. For instance, given that adolescence is a period where some agential capacities seem to develop faster than others, some suggest that we may have to consider the possibility that different aspects of moral agency may come apart (Hartvigsson & Munthe, 2018). This claim may be taken to echo pluralist accounts of moral agency (3.2.3), but likewise seems to reflect a commitment to the idea of distinct aspects or dimensions of moral agency, such as the defendant-claimant distinction suggested in Paper III.

4.1.7 Concluding Remarks: Epistemic Humility and the Possibility of Moral Heterogeneity

Before turning to the nonhuman cases at center stage in this thesis, namely artificial intelligence entities and nonhuman animals, I would like to consider the accuracy desideratum suggested in 3.3 in relation to the mentioned discussions about nonstandard human cases. While I am committed to a naturalistic approach to moral agency and responsibility, I believe that such an approach likewise requires that we are aware of and acknowledge the potential epistemic shortcomings of speculative generalizations.

I find that the discussion of neuropsychiatric and other psychiatric conditions in the philosophical literature on moral responsibility and moral agency tends to make rather quick and simplistic assumptions about highly variable conditions and nontypical forms of functioning to drive philosophical arguments. For example, one may note the absence of nuance in the picture painted of rather large and variable groups of people, such as autistic persons (see Richman & Bidshahri, 2018), persons with intellectual disability (see Carlson 2009), persons with ADHD, or persons suffering from depression, just to mention a few. For instance, the reliance of autistic persons on nontypical ways to know the mental states of others may, depending on the exact underlying condition, be explained by different neurological, cognitive, and behavioral mechanisms and strategies that are not necessarily cases of incapacity (see also Radoilska, 2023).

In addition, there are various (sometimes competing) theories of what the core of autism, as well as many of the other mentioned conditions, actually are. Hence, even assuming that one is confident in a particular account of moral responsibility

and agency, the various factors, possible normative considerations, and scientific uncertainties taken together seem to favor cautious and "messy" rather than certain and "neat" verdicts about the moral agency of agents "with certain impairments" (Jeppsson, 2022b, p. 82; see also Richman & Bidshahri, 2018).

What is more, there is reason to question the assumption that one can learn all there is to know about the subjective or phenomenological dimension of moral agency merely from consulting one's own, and one's peers', experience. This means that one should avoid accepting from the get-go that one can learn whether and why a certain agent is a moral agent or not merely by consulting the feelings and action tendencies one has, or imagines that one might have, in the position of or toward that particular agent.

The risk of relying too heavily on our own *experiential data* as epistemic markers for whether someone is a moral agent is that our moral emotions and action tendencies might, in fact, not (only) be responding to relevant considerations. Our intuitions and judgments may, in many of these cases, also be influenced by irrelevant social factors, such as prejudice and knowledge gaps (see, for example, Jeppsson, 2023). Hence, one should avoid speculative generalizations in favor of empirical evidence.

There is also reason to be skeptical of assumptions about homogeneity regarding the experience and underpinnings of moral agency (see, for example, Westra & Andrews, 2022). I believe that accuracy requires seriously considering the testimony of philosophers and others with first-hand experience regarding the moral psychological phenomenology of this wide range of psychologies, conditions, and disabilities. There is a rich literature of first-person testimonies that seem to support a much more diversified, and nuanced, picture of nonstandard human moral agency. Some of this literature is argued to, for instance, challenge the "popular preconception that people with autism are unable to truly empathize with others due to emotional deficits in a way that affects their moral competence" (Radoilska, 2023, p. 10).[48] Taking seriously such testimonies is paramount for valid assessments (see Stenning, 2020; Radoilska, 2023; Jeppsson, 2021, 2022a).[49]

---

[48] For example, Shoemaker (2015) believes that "those with high-functioning autism tend to experience neither guilt nor pride (or experience them only in rare cases" (2015, p. 171).

[49] See Stenning (2020) for a discussion about autistic life writings (encompassing works from the last three decades) as an important but overlooked potential source of support for autistic moral agency. See Radoilska (2023) for a philosophical account of moral competence that partly builds on first-person testimonies of people with bipolar disorder, autism, and schizophrenia. See Jeppsson (2022a) for a philosopher's first-person account of living and dealing with

Secondly, it is not evident that there is much unity even in the way (typical adult) philosophers think about and react to allegedly nonstandard cases. Philosophical accounts of lived experience with allegedly agency-undermined people are claimed to show that philosophers may often overestimate the extent to which their armchair intuitions and assessments are representative. For example, Kittay (2009), philosopher and mother of disabled daughter Sesha, disagrees strongly with the assumptions on which Singer (1996) and McMahan (2002) base their respective claims about the moral patiency of people with cognitive disabilities.

However, aiming for accuracy and applicability also highlights further methodological issues. Taking a more comprehensive look at empirical data in the psychology of moral appraisal makes apparent a non-negligible discrepancy between the features assumed to make for appropriate and fitting inclusion, excuse, and exemption, and when and why we actually include, excuse or exempt. Recent years have therefore seen an increasing number of philosophical works raising concerns about and calling attention to the undue influence of external conditions that "shape, scaffold, or undermine, agency" (Kennett & Wolfendale, 2019, p. 39; see also Carbonell, 2019; Ciurria, 2023; Hutchison, 2018).

In addition, including or exempting someone as moral agent "can be enormously significant for both legal and social matters" (Richman & Bidshahri, 2018, p. 49). The fact that many nonstandard cases are also part of socially vulnerable groups increases the risk that irrelevant factors influence whether, and to what extent, they are subjected to or exempted from various responsibility responses. Assuming that there are harms or costs to the various ways one is included or exempted as moral agent raises the question whether, and to what extent, the marginalization of some groups is due to their differential position in moral responsibility practices. The social constitution of moral agency in general, and the potential psychological effects of exemptions in particular, are the main supportive arguments employed in favor of the normative claim I develop in Paper IV.

As stated in 3.3.2, however, the objective of valid comparisons points to a tension between descriptive desiderata, in terms of accuracy of the practices as they stand, and prescriptive desiderata, in terms of guidance of when it is (in)appropriate to hold responsible (Argetsinger & Vargas, 2022). I will return to

---

relapsing episodes of psychosis and Jeppsson (2021) for an argument that psychotic phenomena can be intelligible to a greater extent than usually thought, supported by, in part, first-hand experience.

these methodological questions, and discuss a possible solution, in the final discussion (Chapter 6).

4.2 Beyond Human Moral Agency

This section serves to outline the debates on nonhuman moral agency, more specifically the possibility of moral agency in nonhuman animals and artificial intelligence entities. The first nonstandard nonhuman case considered in this thesis is artificial intelligence entities. This is then followed by a similar account and analysis of the second nonstandard nonhuman case considered, namely, nonhuman animals. These cases raise different intuitions and highlight partially distinct questions in relation to moral agency.

A common denominator of these discussions, however, is that they each make explicit problematic assumptions about standard (human) moral agency. Assumptions that undermine the possibility of conducting valid comparative assessments of moral agency. In this section I wish to show the benefits of an empirically informed, theoretically modest, and operationalizable account when asking questions about moral agency in nonhuman, as well as other, nonstandard cases.

4.2.1 Artificial Moral Agency

A detailed overview of the artificial moral agency (hereafter AMA) debate can be found in Paper I. The following section provides a shorter summary of this debate followed by a discussion about some of the themes and issues identified.

4.2.1.1 Linking the Artificial Moral Agency Debate to General Moral Agency Approaches

In Paper I we provide an extensive overview of the AMA debate. Without overly repeating the specifics of that review, we find some central issues related to the idea of machines possessing moral agency. First, one main finding is that much of the debate is locked into a capacity-focused approach, where typical opponents of artificial moral agency hold that moral agency requires subjective mental states and metacognitive capacities of a sort supposedly unlikely to be possessed by machines, while those more open to artificial moral agency instead embrace a set of behavioral capacities. One example of this is the disagreement between followers of Deborah Johnson's (2006) insistence on phenomenal consciousness and those sympathetic to Floridi and Sanders' (2004) idea that a set of behavioral dispositions may suffice (see Paper I for details).

A pragmatic skeptical approach to the consciousness requirement is found in Johansson's (2010) *as-if* approach. While agreeing that moral agency may require subjective mental states in line with the standard view, Johansson stresses the problems raised by indiscriminately holding this requirement against artificial entities. As pointed out by functionalists, human beings attribute relevant mental states, and in turn moral agency using observable features. To be able to retain our ordinary practices of ascribing consciousness and moral agency, the underlying pragmatics of such practices will need to be applied to machines as well. Assuming the possibility of *multiple realizability*, namely, that a mental state, such as pain, can be realized by different physical kinds (artificial and biological) (Putnam, 1967), may then support the attribution of said state to an artificial entity that exhibits the requisite observable features of pain.

Another central divider in the artificial moral agency debate is the question of whether, and in what sense, free will or autonomy is necessary for moral agency. This requirement appears to boil down to the question of *sourcehood,* namely whether artificial entities can be the source of their actions in a way that allows them to be attributed moral authorship (see McKenna & Coates, 2021).

Being eligible for moral responsibility is argued to require being the ultimate source of one's values, decisions, and actions, and consequently that one has come to acquire those features *in the right way*. Because machines and other artificial entities lack source control, for example, in terms of having "authentic goals and values", they can never be moral agents (Hakli & Mäkelä, 2019, p. 269; see also Johnson, 2006).

Floridi and Sanders (2004), among others, resist the source control argument against machine independence by pointing to the fact that such claims do not adequately consider the implication for human moral agency. Like machines, human intentions, decisions, and actions are formed by genes and environmental factors (also see Powers, 2013). Hence, if the possibility of AMA is denied on the basis of design and programming, we may seem to be left with a source control condition too restrictive even for humans (Paper I, sec. 5).

Another central skeptical claim against artificial moral agency is found in the argument that moral agency requires more than mere practical rationality. For an entity to be a moral agent it also needs *moral competence* (Sliwa, 2016; Macnamara, 2015b; see also sec. 3.2.2 and 4.1.1, and Paper I, sec. 4). However, the possibility for artificial moral agency in light of this requirement seems to, once more, depend on whether abilities making up moral competence can be understood in terms of measurable features. Some deny this possibility on the grounds that moral

competence requires features like "phenomenal consciousness, phronesis, or the intuitions required for wide reflective equilibrium" (Purves et al., 2015, p. 11). However, as already mentioned, these features all seem describable in dispositional/functional terms, and other claims to the effect that phenomenal consciousness is necessary for moral agency seem to lead back to the epistemic issue discussed elsewhere (3.3.2 and Paper I, sec. 3).

In addition to the mentioned disagreements on the necessity of standard requirements, several authors suggest alternative approaches to the question of artificial moral agency. For example, different *types* or *degrees* of moral agency or moral responsibility, such as "role responsibility" (Johnson & Powers, 2005, p. 100) or *virtual* moral agency and responsibility (Coeckelbergh 2009, 2010) just to mention a few. Many of these suggestions can be described as attempts to circumvent (some of) the conceptual and practical problems raised in the artificial moral agency debate (see Paper I).

However, many of these alternative suggestions may likewise be considered serious contributions to the general philosophical discussion on moral agency and responsibility. For example, Watson (1996/2004) and Shoemaker (2015) appeal to the complexity and ambiguity of nonstandard human cases to motivate their respective pluralistic suggestions. In a similar sense, considerations of artificial moral agents can be said to make explicit limitations in current concepts and accounts of moral agency and responsibility. In this way, considerations of possible future artificial agents, as well as the emergence and experience of new types of settings, interactions and practices involving such agents, may be important drivers of, and sources of, future philosophical research on moral agency and responsibility. This, if anything, seems to favor a less theory-driven and a more normatively oriented approach.

Such an approach is precisely the suggestion made in Paper I. We argue that the artificial moral agency debate fails to properly address and provide solutions to increasingly pressing practical and normative ethical issues. Instead of assuming from the get-go certain requirements of moral agency and investigating whether these can be met by artificial entities, we argue for a normative turn. In other words, we argue that one should ask whether and how machines *should* be included in human practices where moral agency is normally assumed. Conditions for moral agency may be used in such a discussion as well but should be evaluated in a normative ethical context.

4.2.1.2 Normative and Practice-Focused Approaches to Artificial Moral Agency

In recent years, there have been some explicit normative suggestions to artificial moral agency (Nyholm 2018, 2019; 2020; Johnson & Powers 2005; Verbeek 2011). A common theme is to argue against ascribing moral agency to machines or against designing or using machines in a way that may cause people to see and treat them as moral agents (van Wynsberghe & Robbins, 2019; Hallamaa & Kalliokoski, 2020; Fritz et al., 2020; Boddington, 2021). A common assumption is that seeing and treating artificial entities as moral agents could pose potential harms or risks. These harms or risks are, however, not limited to considerations about safety in relation to a particular application, but rather concern more general worries. One worry is that engaging with artificial entities in moral practices "may change the basic conditions of human action and, therefore, the intrinsic value of human dignity" (Hallamaa & Kalliokoski, 2020, p. 3). Another claim is that due to "the importance of our moral agency" humans "should never relinquish moral agency to machines" (Boddington, 2021, p. 109).

Other normative considerations can be found in pragmatic or instrumental reasons in favor of, or against, including or exempting artificial agents as moral agents. As such, these discussions are similar to some of the pragmatic or instrumentalist suggestions found in the human-centered debates (McGeer, 2019; Vargas, 2013; see Vargas, 2022). For example, some suggest that machines can be related to in a similar way we already consider the behavior of children or pets. Just like the harmful behavior of, for example, a dog is fundamentally the (moral and legal) responsibility of their owner, something similar is argued to be a possible solution to purported *responsibility gaps* posed by AI (Matthias, 2004). The harmful behavior of advanced and autonomous machines can, and should, be the moral responsibility (and legal liability) of human guardians or owners (Köhler et al., 2017; Schaerer et al., 2009).[50]

A similar suggestion is made by Tigard (2021) who argues that the resources available from an objective stance, such as control and training, seem to suffice for a wide range of possible problematic scenarios posed by autonomous and advanced artificial entities. What is more, Tigard (2021) among others seems to think that once human users (or victims) consider and learn that some (perceived) harm or injury was, in fact, due to a machine and not a human, they will undoubtedly switch to the objective stance. In this way, the alleged issues posed

[50] See Johnson and Verdicchio (2018) for a critical discussion of animal-robot analogies.

by responsibility gaps, originally put forth by Matthias (2004), are argued to be exaggerated and easily solved by making evident the artificial nature of the entity.

While I am sympathetic to Tigard's (2021) practice-focused approach to questions about responsibility gaps, the suggested solution does not seem to consider recent empirical data. In particular, Tigard appears to underestimate the tendency of humans to readily anthropomorphize and socially engage with artificial entities. Mind attribution, for instance, is not something over which one has complete conscious control. On the contrary, human cognition is very sensitive to cues, such as "the presence of eyes and directed gaze, goal-directed motion, and self- propelled motion" (Epley & Waytz, 2010, p. 521; see also Terada et al., 2007) as well as social interactive behavior (Hortensius & Cross, 2018), all of which increase the probability of mind attribution, anthropomorphizing of, and social interaction with artificial entities (Martini et al., 2015; Saltik et al., 2021; Hortensius & Cross, 2018; Thellman et al., 2022).

What is more, there are various studies purported to show that humans already judge machines as blameworthy given the presence of certain features, such as perceived intentional harm together with perceived phenomenal states (Sullivan & Fosso Wamba, 2022; Bigman & Gray, 2018). Other studies even suggest that humans morally appraise AI in ways reflecting a pluralistic understanding of moral responsibility and agency. That is, people do not only morally appraise the actions of AI, but also the character of such entities. Hence, humans ascribe responsibility to artificial agents both in a sense that looks like accountability and in a sense that looks like attributability (Lima et al., 2021; Gamez et al., 2020).

And while the blaming responses of humans to (current) AI may differ in some respects from those usually seen in human-human interactions (Malle et al., 2019), this may simply be a function of how socially embedded a particular agent is. If we move to consider a plausible future scenario in which robotic pets, companions and caregivers, virtual assistants, and so forth, are improved in terms of conversational ability, emotional intelligence, and emotional expressivity, it seems imprudent to deny the possibility that humans may come to see and treat some of these agents as moral agents.

I believe that discussions regarding the possibility or risk of seeing and treating socially capable artificial entities as moral agents need to take seriously the fact that perception and attribution of moral agency do not solely depend on explicit and conscious processes. Hence, even "once the facts are known" (Tigard, 2021, p. 594), the appearance and behavior of an artificial agent may very well still elicit something akin to the participant stance in human users. Hence, I am skeptical of

the utility of labeling an AI as "artificial", or designing it to explicitly disclose its artificial nature, if the entity at the same time is devised to function in ways that elicit mind attribution and encourage emotionally engaged social interaction.[51]

For example, if one wants to avoid the ascription of moral agency to artificial entities, one needs to consider the inherently social and emotional nature of such ascription. Rather than assuming that the participant and objective stances are adopted on basis of conscious reflection, a normative and practice-focused approach requires us to learn about the actual features and contexts in which humans adopt and switch between these perspectives.[52]

Another type of normatively driven view on artificial moral agency is found in arguments about the moral superiority of artificial intelligence entities. According to this view, artificial intelligence entities have much better prospects of behaving morally than humans. This is so, proponents claim, precisely because artificial agents lack certain human features, such as emotions and other phenomenal states. While this position may strike some as absurd, weaker varieties of this claim are, in a sense, already used to motivate the development and use of AI applications in procedures such as "hiring, lay-off, university admission, and loan approval decisions" (Claudy et al., 2022, p. 4). The use of such applications is often supported by the promise that they will enhance "the accuracy, consistency, and incorruptibility from the social influence of many decision procedures." (2022, p. 7).

Similar arguments are also put forth in favor of the design and use of lethal autonomous weapons systems (or LAWs, see, for example, Müller, 2016; Umbrello et al., 2020), such as unmanned combat aerial vehicles (see Austin, 2010). Human soldiers are vulnerable to fatigue and emotional distress both of which undermine performance in terms of, among other things, the accuracy of shots (Burke et al. 2007; Nibbeling et al. 2014). At the same time, LAWs are artificial and thus unaffected by circumstances that normally cause emotional or physical stress. This is claimed to make them ethically superior to humans in combat (see, for example, Umbrello et al., 2020).[53]

---

51 Gunkel (2018a) argues that we need to take seriously that appearances may often trump "revelation" (2018a, p. 18).

52 And this may very well also be the case for a range of features or statuses, like phenomenal consciousness discussed earlier. As such, we need to learn about, and be informed by, the features or circumstances that tend to elicit a specific perception or attribution and particular attitudes and behaviors (see, for example, Epley & Waytz, 2010).

53 See Klincewicz (2015) for a counterargument that appeals to the vulnerability of computerized combat systems to hacking.

With regard to the question of moral agency, Gips (1995) claims that while "not many human beings live their lives flawlessly as moral saints … a robot could" (Gips 1995, pp. 249-250). Similarly, Dietrich (2001) argues that the emotions and limitations of humans make us "genetically hardwired to be immoral" and that we should therefore replace ourselves with artificial agents, whose rationality and impartiality, make them "a vast improvement over us" (2001, p. 326). Similar ideas are put forth by various authors (see, for example, Arkin, 2009). The idea seems to be that some aspects of the psychological make-up of humans limit our ability to be impartial, consistent, and unbiased in ways important for moral decision-making and behavior.

These arguments thus seem to assume that some emotions, desires, and attitudes stand in the way of moral reasoning and decision-making. As such, the superiority position connects to the general question of whether rationality and practical agency make up the core feature of moral agency, or whether such agency requires phenomenal states and processes, such as sympathy, feelings of obligation, guilt, etc. As mentioned earlier, these themes and issues are central to discussions about some nonstandard human cases assumed to lack empathy, sympathy, moral sense, or moral competence (for instance in the case of the psychopath, 4.1.1).

However, as mentioned in the previous chapter, empirical studies of human moral psychology support the view that emotions are fundamental to moral decision-making (Haidt, 2001, 2012; Prinz, 2007; Greene et al., 2001; Koenigs et al., 2007; Damasio et al., 1991; Nichols, 2004). For instance, emotional empathy, feeling how another feels, appears to be central to moral judgment (Aaltola, 2014; de Waal, 2010; Decety & Cowell, 2014; Kauppinen, 2017; Shaw et al., 1994; Zaki, 2018).

What is more, assuming that our ascriptions of moral responsibility have quality of will as their target, moral agency appears to presuppose subjective attitudes. Resentment and indignation are, after all, reactions to perceived ill will or indifference. We are prone to these responses because we care about the attitudes of others toward ourselves (and others). Hence, if an agent does not have a subjective viewpoint in the first place, they would not be capable of harboring attitudes to which our reactions of moral responsibility are appropriately sensitive.

Against this background, it does not seem that human (or animal) emotional nature stands in the way of moral agency, but rather that emotions make such agency possible in the first place. Rather than impartiality and *coolness*, sensitivity and responsiveness to moral considerations seems to rely very much on modes

and processes involving subjective perspective and emotional engagement.[54] Hence, if current (or possible future) machines lack subjective emotions, desires, and attitudes, they would not have the skills required for being responsive to moral considerations.[55] Nor would they be open to moral appraisal.

4.2.1.3 Concluding Remarks: Measurable Standards and Moral Patiency

To conclude, the debate on artificial moral agency highlights several interesting themes. First, discussions about autonomy, control and sourcehood make apparent possible incompatibilist intuitions prevalent in these debates. Despite being created and designed, artificial entities could come to have the ability to change, learn and adapt flexibly and even self-modify in response to environmental challenges, not unlike humans (and other animals). Requiring something more metaphysically heavy in terms of being the ultimate source of one's actions therefore runs the risk of undermining the possibility of human moral agency as well.

Secondly, the disagreement about phenomenal consciousness makes explicit that there is wide-ranging skepticism about the possibility of phenomenally conscious (or sentient) artificial entities. However, this assumption serves to highlight general epistemic and pragmatic questions that are seldom addressed in the general moral agency and responsibility debate, in particular the "problem of other minds": we simply cannot have direct access to the possible subjective states of another, not even another human.

As such, we seem to have to recognize that attributing phenomenal consciousness or particular mental states, be it in everyday encounters or in scientific research, rely on observable and measurable features, such as observations of behavior and neuroimaging of brain states (see Avramides, 2020). In practice, attributions of consciousness and other mental states and processes to an entity depend and rely on how the agent interacts with, and respond to, others.

In this way, the sourcehood and consciousness discussions not only highlight issues about over-intellectualizing moral agency and treating it as an intra-individual competence. These discussions also make explicit that some assumed requirements held against artificial moral agency, are in fact non-verifiable. That is, consciousness, sourcehood, et cetera are often conceived in such vague ways that they do not allow for systematic assessments. The problem, then, is not only

---

54 As we will see, this is a line of thought that is prominent also among proponents of animal moral agency.

55 It is important to note that this argument does not necessarily assume an intra-individual conception of emotions or attitudes and thus of moral agency. See Paper I, sec. 4 and 7.

that the mentioned debates are characterized by anhropofabulation (Buckner, 2021), but also that skeptics of artificial moral agency often assume standards that are impossible to observe or measure empirically.

But given that moral agency is manifested socially, assessments of moral agency and responsibility in the case of artificial intelligence agents should be conceived as a social-normative competence. While I will not be able to provide a robust ready-to-use operationalized definition here, I am certain that the prospects for arriving at tangible criteria assuming a practice-focused approach are good.

Consider, for instance, the explicitly interactive moral responsibility *face* of answerability (Shoemaker, 2015, Ch. 2). Moral agency in terms of answerability requires being capable of providing answers for one's actions by morally responding to other agents' "demands for justification" (A. M. Smith, 2012, p. 578). Assuming this understanding of moral responsibility, some current machines and programs, such as advanced chatbots, appear to be possible candidates in the sense of being able to provide reasons for their actions or omissions. On the other hand, even this specific form of moral agency is argued to require phenomenal states and processes - namely, the guiding sentiment of agential regret (Shoemaker, 2015, Ch. 2).[56]

The question of emotions or phenomenal consciousness also links the artificial moral agency discussion to debates about the possibility of artificial intelligence moral patients. Given that sentience is typically considered a basic condition of moral patiency, the possibility of subjective experience in machines raises ethical questions. Can, and should, artificial entities be afforded moral patiency (Gunkel, 2018b; Danaher, 2020)? If so, what are the practical and normative implications? For example, on some versions of utilitarianism the prospect of happy artificial entities may provide us with strong reasons to create such entities.

However, according to a less optimistic approach, the prospect of a new class of sentient beings may instead be seen as providing a strong reason to abstain from developing such entities. Given the historical *track record* of humans with respect to, for example, catching, confining, experimenting on, and slaughtering sentient nonhuman animals, we simply do not seem morally fit to co-exist benignly with an additional class of sentient creatures (Behdadi, 2019). This worry finds additional support in the fact that the primary purpose of artificial entities is to serve and assist humans. As such, the implications of artificial intelligence moral patients may look even worse (see also MacLennan, 2013; Metzinger, 2013). In

[56] See also Coeckelbergh (2020b) for an argument stating that the case for *explainability* in AI can and should be approached and justified in terms of human answerability practices.

addition to sentience-based considerations of moral patiency, the possibility of artificial moral agency can itself actualize questions about such patiency. I will return to this question in the next chapter (5.2.3).

4.2.2 Nonhuman Animal Moral Agency

The consideration of nonhuman animal moral agency links to complex research topics in nonhuman animal psychology and cognitive science, as well as basic moral psychology. Scholars are increasingly asking whether nonhuman animals can feel, think, behave, or act in ways that can be described or assessed as *moral* (see, for example, Rowlands, 2012; S. Fitzpatrick, 2017; Delon, in press; Clement, 2013).[57]

While these discussions do not necessarily represent or involve opposing accounts of or verdicts about (animal) moral agency, they are important to the extent that they are relevant to considerations regarding accuracy and applicability. In this thesis, this is especially evident in Paper II and Paper III.

4.2.2.1 The Metacognitive Paradigm: Skeptical Views on Animal Moral Agency

Although nonhuman animals have been reported to exhibit seemingly moral, prosocial (Decety et al., 2016), empathetic, or virtuous behaviors or emotions like compassion, concern, grief, equity-aversion, care or altruism (de Waal, 2014; Vincent et al., 2018; Delon, in press; Andrews, 2020b; Westra & Andrews, 2022; Monsó & Andrews, 2022), various authors maintain that this does not indicate moral agency. For instance, Dixon (1995, 2008) denies that a dog who pulls a child from a fire can be ascribed moral motivation and thus moral agency. Instead, she claims, it is more likely that "the dog is made anxious by the cries for help and only wishes to stop these sounds by the most expedient method—removing the child from the burning building" (1995, p. 40)

Animal moral agency skeptics, like Korsgaard (2006; 2010), Ayala (2010), Musschenga (2015), Kitcher (2011), and Dixon, (1995, 2008) typically defend their position by appeal to a metacognitive conception of moral agency. The basic

---

[57] A predominant area of debate concerns the extent to which moral (or more widely, normative) cognition, thought and behavior may be shared beyond the human species (Andrews, 2020b; Flack & de Waal, 2000; de Waal, 2006, 2014; von Rohr et al., 2011; Danón, 2019; Kagan, 2000; Bernstein 2000). This debate is sometimes referred to as the continuist-discontinuist disagreement (Cova, 2013). Another prominent area of inquiry concerns what normative cognition consists of. For instance, what, if anything, defines normative, as opposed to, other types of thoughts, feelings, motivations, and behaviors (Vincent et al., 2018; Lorini, 2022; Andrews, 2020b; de Waal, 2014; Danón, 2019)?

assumption is that higher-order, or metacognitive states or processes, like reflection, evaluation, and self-consciousness, are necessary for moral agency.[58] A common reason put forth for this position is that moral agents act *from* moral rules or principles and not merely in accordance with such rules or principles.[59] For a being to be a moral agent her actions need to arise in the right way. They need to, strictly speaking, be her *own*. In this sense, these skeptics assume that conscious deliberation is what makes an entity the proper source or author of her own conduct.[60] As we will see later, those who are more optimistic to the possibility of animal moral agency tend to precisely question this claim.

According to some skeptics to animal moral agency, conscious reflection is necessary because it enables the agent to rise above their immediate emotions or impulses and therefore grants them the kind of control necessary for moral agency. For instance, Korsgaard (2004) writes: "as rational beings we [humans] are conscious of the principles on which we are inclined to act. Because of this, we have the ability to ask ourselves whether we should act in the way we are instinctively inclined to. We can say to ourselves: "I am inclined to do act-A for the sake of end-E. But should I?"" (2004, pp. 148-9).

Note that the appeal to emotion as undermining control rarely (if ever) figures in the artificial moral agency debate, where the prevalence or possibility of phenomenal states, such as emotions, in artificial entities is, as mentioned, far from a given. On the other hand, the assumed issue presented by natural impulses or "instincts" of animals appears to have some parallels in the artificial moral agency debate. For example, the idea that programming renders artificial entities inflexible and determined looks similar to the way some view animals as bound or determined by their biology. For instance, Korsgaard (2006) believes that because nonhuman animals are merely moved by their affective states, such as emotions and desires, they are but "in Harry Frankfurt's phrase, wanton" (2006, p. 102).

Another skeptic, Ayala (2010), argues that conscious deliberation is necessary for morality, because it makes one capable of *anticipating* and *imagining* possible effects and consequences of actions and omissions. This capacity to "establish the connection between means and ends" (2010, p. 9018) is fundamental to morality

---

58 See also Monsó and Andrews (2022) for different uses of the concept metacognition in relation to animal morality.

59 Dixon (2008) is skeptical of animal moral agency and defends a view where emotions are central to such agency. However, praiseworthiness requires that emotions are accompanied by certain explicit moral beliefs.

60 See Sapontzis (1980) for a similar account of the traditional reason-based arguments against animal morality.

because it is necessary for two other conditions of moral behavior: namely, the ability to make value judgments, and the ability to rationally choose between different courses of actions. Because nonhuman animals lack "man's eminent intellectual abilities" (Ayala, 2010, p. 1915) to imagine and consider future scenarios and actions, they cannot act morally. Instead, they are wholly determined by their desires and instincts, unable to critically assess their motivations and therefore unable to, in any real sense, *choose* how to act. As such, they lack the kind of higher-order thinking that underpins the control necessary for moral agency.

Metacognitive abilities, like reflection, are also thought to be necessary because they underlie or make possible moral competence or moral knowledge. For example, Ayala (2010) argues that conscious deliberation is necessary for a being to have a *moral sense,* namely, the ability to perceive and judge the moral value of different outcomes. Merely living or behaving in a way that is consistent with some moral norm or promotes *the good* is thus not sufficient for being a moral agent. These claims echo some of the competence-based views found in the general moral responsibility and agency debate. For instance, the position that moral agency requires some additional understanding, skill, or competence, such as competence to engage in the type of moral conversations fundamental to moral responsibility practices (see, for example, McKenna, 2012).

However, requiring conscious reflection for moral competence seems to go against sentimentalist views on (human) moral sensitivity and responsiveness. As mentioned, having emotions, desires, and attitudes is by many argued to be central to moral agency. For instance, many argue that moral motivation necessarily involves phenomenal states, like emotions and many also claim that moral knowledge requires experiential capacities (see previous chapter, 3.2.2, and this chapter, 4.1.1, 4.2.1).

A developed and rather nuanced skeptical argument regarding animal moral agency has been put forth by Musschenga (2015), who acknowledges that many animals are capable of *certain forms* of self-control, namely inhibition of habitual behavior. Because of this, some animals can adapt their behavior in light of situational circumstances and, in this sense, respond to reasons. As such, some animals are what one may call "reason trackers" (2015, p. 53). However, Musschenga maintains that reason tracking is not sufficient for the deliberate direct control required for the type of sourcehood involved in moral agency. Following Fischer and Ravizza (1998), he argues that moral agency additionally requires reason responsiveness, the ability to respond differently when presented with different

reasons.[61] This, in turn, is made possible by the capacities for deliberate reflection and evaluation typically found in humans.

To reiterate, the metacognitive argument against the possibility of animal moral agency holds that conscious reflection and evaluation is necessary for moral agency in virtue of constituting or underlying capacities for control and moral competence. Metacognition is necessary for control because only a creature capable of rising above its immediate feelings and inclinations can *choose to act*, as opposed to merely behave, and thus be held accountable for its behavior. Metacognition, the argument goes, is necessary for moral competence because only a being that understands and recognizes moral obligations and the morally relevant features of a situation can act for the *right kind of reasons* and be attributed the moral qualities of their behavior.

In this way, the main skeptical points in the artificial and animal moral agency debates appear to go in opposite directions. While skeptics of artificial moral agency underline the importance of phenomenal states, such as feelings and emotions, for being sensitive and responsive to moral considerations, skeptics of animal moral agency emphasize traditional cognitive features, such as rising above one's feelings and emotions. These differences are clearly, at least in part, due to different intuitions and assumptions about phenomenal consciousness in each case. Philosophers attribute sentience to (many) nonhuman animals but are skeptical about the prospect of sentient artificial entities.

4.2.2.2 The First-Order Paradigm: Proponents of Animal Moral Agency

Needless to say, several authors in the animal moral agency debate oppose the arguments put forth in favor of the metacognitive stance, as well as the often-corresponding skeptical conclusion regarding morality or moral agency in animals. Instead of assuming conscious reflection and evaluation to be necessary, the proponents of animal moral agency defend accounts where attribution of rightness/wrongness, goodness/badness or virtue/vice require only certain first-order mental states or processes.

However, it is worth stating that, like the more skeptical accounts above, many authors in this latter camp are still hesitant to ascribe full-fledged moral agency to nonhuman animals. Instead, their optimistic verdicts are often restricted to less substantial, or alternative, conceptions of moral agency, which do not imply moral responsibility. Alternatively, the few authors who do claim that nonhuman animals

[61] Musschenga (2015) adopts the distinction between reason trackers and reason responders from K. Jones (2003).

can be morally responsible in a full-fledged or substantial sense are still hesitant about whether humans can, or should, hold other animals responsible.

As we will see, most of the optimist accounts can be characterized as belonging to the quality of will or moral responsiveness tradition (3.2 and 3.3.2). As such, they often appeal to sentimentalist accounts of moral motivation and behavior (see Kauppinen, 2022) that emphasize first-order states and processes, such as sympathy or concern, as opposed to metacognitive states and processes, such as reflection or conscious deliberation (see Clement, 2013). In addition, many proponents of animal moral agency appear lean toward conceptions of moral agency that understand it in terms of eligibility for aretaic appraisals or judgments about character, rather than in terms of accountability or answerability (Shoemaker, 2015).

The first type of account in favor of animal moral agency is found in minimalist attributionist or virtue ethical accounts of moral agency and responsibility (Clark 1984, 1985; Sapontzis 1980, 1987; Shapiro 2006; Burgis, 2018; Waller, 1997; Wrage, 2022). According to this view, if a being acts from apparent virtuous character traits or emotional states (like compassion), or expresses intentions or motivations that are virtuous, this is sufficient for the attribution of moral responsibility in the sense of virtue or aretaic praise to that being or their actions.

Contrary to the metacognitivists, attributability-centered arguments in favor of animal moral agency therefore deny the relevance of conscious reflection. Waller (1997) argues that the idea that acts either result from conscious deliberation, or they are akin to mindless, involuntary stimuli-responses is false. There is, he claims, ample room between these two extremes. Acting virtuously or morally does, Waller admits, require that one acts for the right reasons. But reason as such is not required for moral behavior. Hence, while nonhuman animals may not be able to explicitly reflect on or evaluate about principles or theories, many of them are argued to have what is required for moral appraisal. The kind of appraisal discussed in these accounts, however, seems to be restricted to only the attributability sense of moral responsibility (Shoemaker, 2015).

A similar type of claim in favor of animal moral agency is found in evolutionary arguments. The core claim here is that morality, in terms of moral competence and practices of moral responsibility, is widespread in the animal kingdom. For example, de Waal (1996, 2006; Flack & de Waal, 2000) argues that it is a mistake to assume that human morality is, but a thin veneer situated on top of a brutish animal nature. To the contrary, he argues that the basic building blocks of morality are emotional and shared with many other species, such as apes. It is therefore a

mistake, de Waal argues, to view morality as a uniquely human competence or capacity to rise above one's desires or appetites. Morality is not at odds with our animal nature, but an extension of it.[62] In a similar sense, Bekoff and Pierce (2009) define morality as "a suite of interrelated other-regarding behaviors that cultivate and regulate complex interactions within social groups" (2009, p. 7). The authors argue that understanding morality in terms of its adaptive value makes explicit that it is found in various social animal species beyond our own, such as canids (Bekoff & Pierce, 2009; see also Paper II).[63]

Seeing morality in terms of its function(s) is thus assumed to support the view that "inclinations-caring morality" is the foundation of moral behavior, while "rational-duty morality" is a complementary (human) extension (Waller, 1997, p. 354; see also de Waal, 2006, 2014). Humans are certainly capable of providing verbal accounts of their intent, and this seems to require more complex capacities. But these reflective and linguistic capacities are not essential to morality (as discussed in Paper II and III). Rather, our commitment to care and trust is the *foundation* of moral deliberation, rules, and duties. In this way, reason-based morality does not transcend a brutish past but may serve to enhance and extend the sympathetic biology already present (de Waal, 2006, 2014; Waller, 1997).

A third prominent strategy employed by the optimist camp is to point to intuitions about everyday human behaviors. For instance, Cova (2013) claims that metacognitive arguments against animal moral agency fail to show that explicit evaluation or judgment is necessary for moral agency, since "[r]eflectivism … seems to discredit every good action that would look like a 'moral reflex' … it also discredits actions that come from emotional reactions and are not mediated by moral reasoning" (Cova 2013. p. 123).

Moreover, "[r]eflectivism seems to lead to the conclusion that, when a friend or a parent helps us, he is all the more praiseworthy for helping us because it was the right thing to do" (Cova, 2013, p. 123). However, people who are kind, helpful or generous are generally not considered less worthy of praise than those who act kindly or generously out of a sense of duty. To the contrary, "a sense of duty is

---

62 De Waal claims that some nonhuman animals, such as apes, have *proto-morality* in terms of possessing the basic building blocks of human moral emotions and behavior. Chimpanzees, for instance, are keen on preserving harmony in the group by "reconciling after conflict, protesting against unequal divisions, and breaking up fights amongst others" (2014, p. 200), and therefore have the normative character, or *ought*, fundamental to morality.

63 For instance, a popular view about (human) moral responsibility is that it should be understood as "a general practice of prosocial influence" (Vargas, 2022, p. 3; see also Schlick,1939/1966 and Dennett, 1984).

commonly considered of secondary value, something which those who unfortunately lack virtuous dispositions must fall back on to do good" (Sapontzis, 1980, p. 51; see also Cova, 2013). In this way, consideration of our responses and intuitions to cases involving humans is assumed to show that metacognitive requirements for moral agency would set the bar too high to account for and justify our everyday intuitions and responses concerning moral responsibility. In consequence, these requirements run the risk of excluding ordinary humans from moral responsibility.

For similar reasons, Andrews and Gruen (2014) suggest an alternative way of approaching questions about nonhuman animal morality: "Once we are able to look past the most salient examples of human morality, we find that moral behaviour and thought is a thread that runs through our daily activities, from the micro-ethics involved in coordinating daily behaviours like driving a car down a crowded street … to the sharing of someone's joy in getting a new job or a paper published. If we ignore these sorts of moral actions, we are overintellectualizing human morality" (2014, p. 194).

For example, when we spontaneously respond to another's distress by coming to their aid, we do not rely on conscious reflection, but on non-reflective care. Similarly, when a mother animal responds to her young's calls of distress and comes to their aid, she is acting from affection for her child. Similarly, most of human moral behavior does not rely on reasoning but is still intentional and responsive to circumstances. Being motivated by loving care for another in this way is sufficient for acting morally and for manifesting one's quality of will (Waller, 1997; Ferrin, 2019; Delon, in press; see Arpaly, 2002).

The possibility of applying instinctual or other "simple" explanations does not, some argue, undermine the morality of animal parental care. This is because the motivation of such behavior is not merely to alleviate the distress caused by the infant's cry or to pass on one's genes (Waller, 1997; Wrage, 2022). Applying Monsó's (2017) account of *Minimal Moral Empathy*, Wrage (2022) argues that animal parental care cannot be explained away in terms of non-moral motivation, since the inclination to act from care or concern for others is not only motivated by the detection of distress in another individual but *aimed at alleviating the distress in the other*. Hence, reductive explanations (such as appeals to mere emotional contagion) do not seem to explain why the mother comes to her young's aid rather than just distancing herself or killing the child.

Second, similar reductive explanations appear to be readily available also in the case of much of human moral behavior. After all, human altruistic behavior has

also been challenged by arguments from psychological egoism. Even if we think that we act on other-regarding reasons, we may be deceiving ourselves (Batson et al., 2011; Shaver, 2023). Third, ultimate or evolutionary explanations of behavior, like the fitness enhancing effects of caring for offspring, do not stand in conflict with having or acting on "genuine selfless concern" (Waller, 1997, footnote 1).

Hence, metacognitive based skepticism to animal morality is argued to put unwarranted focus on "the most rarefied and linguistically mediated" aspects of human moral cognition and behavior (Andrews & Gruen, 2014, p. 209). Excluding everyday examples, which comprise the vast majority of human moral thinking and behavior, is a serious failure on part of the metacognitive position. This position disregards the fact that also in the case of humans, moral thinking and behavior is mainly driven by non-volitional, emotional, and unconscious processes (see Chapter 3, sec. 3.3.2, and this chapter, sec. 4.2.1). What is more, the cool, detached, "ethical point of view" upheld by metacognitivists is far from the actual attached, warm, and "entangled nature" of empathy, care, and prosociality (Andrews & Gruen, 2014, pp. 207-8). As such, the metacognitivist position seems guilty of anthropofabulation by assuming an "exaggerated sense of typical human performance" (Buckner, 2013, p. 853) as the baseline for comparisons with other species.

These arguments echo, and sometimes explicitly refer to, sentimentalist quality of will or moral responsiveness accounts (Arpaly, 2002; Markovits, 2010). And the claim that social engagement and emotion, rather than cool impartiality, is central for moral agency aligns with practice-focused approaches to moral responsibility and agency (Chapter 3). What is more, if sensitivity and responsiveness to moral considerations is made possible by emotions, non-feeling machines would seem to be disqualified. Thus, the mentioned argument also makes explicit a possible challenge to some of the superiority-based arguments used in favor of artificial moral agency mentioned earlier (sec. 4.2.1).

Another optimist argument in favor of animal moral agency is found in explicitly pluralistic suggestions. The basic idea here is that some of the skepticism to, and disagreement about, animal moral agency may be due to a conflation between two distinct senses of moral agency (Cova, 2013; Rowlands, 2012). Pluralistic accounts propose a distinction between being a moral agent in virtue of

having certain first-order psychological states and processes, and moral agency in the sense of certain higher-order psychological states and processes.[64]

A particularly developed pluralistic suggestion is put forth by Rowlands (2012), whose central thesis is that moral motivation and moral judgment are distinct. He argues that since many animals are motivated to act on moral reasons, externally construed, the moral quality of their behavior is correctly attributed to them. They are, in his terminology, *moral subjects*. Animals can be responsive to moral considerations in virtue of having morally laden emotions, that is, emotions that track morally relevant features as part of their content. Moral subjects track moral considerations by being (reliably) capable of perceiving and emotionally responding to (at least some) good- and bad-making features of situations. For example, a being is a moral subject in virtue of perceiving, say, distress in a conspecific, and responding to it by feeling sad and compelled to alleviate it (Rowlands, 2012, Ch. 5 and 9).

None of this therefore requires that the animals entertain, or are capable of entertaining, evaluative propositions, such as moral judgments. It is sufficient that they merely experience the emotion and are moved by it in a reliable manner. Similar to other accounts defending animal moral agency, Rowland's account fits with the quality of will category of views of human moral agency, which conceive of moral responsiveness as responsiveness de re rather than de dicto (see previous chapter, 3.3.2; Arpaly, 2002; Markovits, 2010).

Rowlands (2012) further defends his first-order quality if will conception of morality by stating that metacognition, in terms of reflection and scrutiny, "is not the sort of thing that can confer control over motivations. To suppose that it is would be to fall victim to a version of the fallacy I have labeled the *miracle-of-the-meta*: the mistake of thinking that something magical happens in the transition from first order to metalevel" (2012 p. 183). In this way, Rowlands argument echoes the *hierarchical problem* raised against Frankfurt's (1971) identification account of sourcehood mentioned in Chapter 2 (2.2.1).

However, Rowlands, as most other proponents of animal morality, still agree with the metacognitivists on an important matter, namely, that fully-fledged moral agency in the sense of being eligible for ascriptions of moral responsibility through

---

64 Cova (2013) writes: "First, one can be a moral being because one is morally responsible of (some of) his action: thus, one is a moral being in the sense of being a *moral agent*. Then, one can be a moral being in the sense that one is able to judge whether something is right or wrong: in this sense, one is a moral being in the sense of being a *moral judge*." (Cova, 2013, p. 118). This echoes David Hume's (1777/1975) distinction between compassion and moral evaluation (see also Beauchamp, 1999).

being held responsible or blamed, requires more than the kind of moral responsiveness made possible by first-order states and processes. For a being to be a moral agent in this fuller sense, she also needs to have reflective capacities, according to Rowlands. Hence, Rowlands, as well as most other proponents of animal morality believe that the inability for "critical rational scrutiny" (2012, p. 85) in nonhuman animals renders them improper targets or recipients of moralized blame or praise, or accountability in the retrospective sense (Rowlands, 2012; Shapiro, 2006; Wrage, 2022; Waller, 1997).

In addition, other arguments advocate distinguishing between fuller and lesser forms of morality. For example, de Waal (2006) thinks that self-reflection and logical reasoning are examples of human capacities that enable explicit and universal rules and, by extension, full-fledged moral agency.[65] And, Sapontzis (1980) claims that human beings are moral in a further sense because we are capable of dedicating ourselves "to ideal ways of life" (1980, p. 50).

However, some optimists forward reasons against holding other animals morally responsible in practice, while at the same time maintaining that nonhuman animals can be fully fledged moral agents. For example, Cova (2013) argues that many nonhuman animals are moral agents, but concedes that creatures lacking capacities for conscious reflection are exempted from *punishment*, because "we want the people we punish to understand why they are punished" (2013, footnote 12).[66] In a similar sense, Ferrin thinks it may be the case that "the communication barrier and lack of overlapping social context" (2019, p. 137) stands in the way of holding other animals responsible.[67] In addition to these normative and pragmatic arguments against interspecies accountability practices, Bekoff and Pierce argue that because moral norms and their enforcement are specific to the particular species or communities that have evolved them, "animals are moral agents within the limited context of their own communities." (2009, p. 144).

However, I think that these authors are too quick to reject the possibility and justifiability of interspecies moral responsibility practices. The persistence of this type of pessimism is, I believe, partially due to an overly restricted understanding of moral agency (see 4.3). In Paper III, I suggest that a more differentiated conception of moral agency highlights that participation in moral responsibility

---

65 Similarly, several others have made continuist claims while arguing that human morality is exceptional or unique (Prinz, 2014; Haidt, 2012; Joyce, 2006; Kitcher, 2006, 2011; von Rohr et al., 2011; see S. Fitzpatrick, 2017).

66 See also Arpaly (2002).

67 See also Borchert and Dewey (2023) for an argument in favor of an asymmetry between nonhuman animal praiseworthiness and blameworthiness.

practices involves more than being (held) responsible. We already inhabit some interspecies contexts and seem to have the communicatory resources to morally engage with at least some nonhuman animals. While I do not focus on the possibility of humans holding nonhuman animals responsible, I argue that other animals can, and sometimes do, *hold humans responsible*.

4.2.2.3 Concluding Remarks: Accuracy of Scope and Normative Considerations

In conclusion, the animal moral agency debate highlights some central themes or areas of interest. One prominent disagreement regards the relevance of metacognition, such as reflection, for control and moral competence. Is metacognition, in the sense of, for instance., reflection and deliberation, necessary for moral agency? Or is it sufficient that one is able and inclined to feel, and be moved by, moral emotions or considerations?

In addition, many of the disagreements in the animal moral agency debate seem to revolve around differing conceptions of morality or moral agency. Are moral agents the kind of entities who can be praise- or blameworthy? Or can moral agency be attributed as soon as someone is fit for assessments or evaluations of morality or moral evaluations, say, in terms of attributability (such as wrongdoing or virtuousness)? Many of the optimist arguments limit their conclusions to notions that belong to the latter conception or claim that animals can only be participants in species-specific responsibility practices. In this way even the optimists tend to refrain from asserting the possibility of human-animal interspecies moral responsibility practices.

Hence, even assumed proponents of animal moral agency seem to roughly adhere to the view that morality or moral agency can be understood in two senses: being eligible for having one's motivation and behavior appropriately described and appraised as *moral*, on the one hand, and being eligible for assessments, reactions, and ascriptions of or moral responsibility on the other. This distinction, in turn, can be approached from two directions, resulting in two possible interpretations. First, one can view the first-order and second-order positions on moral agency as, in some sense, tracking two types of eligibilities for moral responsibility. This is similar to Shoemaker's (2015) distinction between being an appropriate target of attributability appraisals, on the one hand, and being an appropriate target of the *social* faces of responsibility (answerability and accountability), on the other.[68] Another possible interpretation, however, is put forth by those who claim

[68] Note, however, that I am not saying that Shoemaker takes his account to imply that first-order/non-reflective states and processes are sufficient for responsibility as attributability.

that this debate is to great extent focused on disagreements where *morality* is conceived of as a "psychological natural kind" (S. Fitzpatrick, 2017, p. 1155).[69] Hence, most apparent disagreements in the animal moral agency debate, it is argued, do not involve opposing opinions on moral agency per se but rather opposing views about the extent to which the allegedly moral behavior or moral psychology of humans are continuous with features found among other animals as well as different positions on moral terminology (S. Fitzpatrick, 2017). As such, many of the arguments put forth in this debate are determined to be non-substantial.

If one were to follow S. Fitzpatrick's verdict about the animal moral agency debate, one may instead say that many of the first-order conceptions of morality or moral agency fail to address the same questions as the positions of many of the skeptics. As mentioned, failing to engage in the same project is something Munthe and I point out as something that characterizes the artificial moral agency debate. However, with regard to nonhuman animals, I believe that S. Fitzpatrick is too quick in his assessment. The claim that moral agency requires being moved to act in response to some qualified type of input, such as distress, and the claim that moral agency requires reflective self-government, are not necessarily answers to completely different questions. Both can be understood as accounts of moral agency by assuming certain conceptually substantial conditions.

A central take-away from the animal moral agency debate, I believe, is that there are obvious accuracy-related problems associated with setting the bar for moral agency too high in terms of requiring occurrent and de dicto moral awareness (see 3.3). If we deny nonhuman animals moral agency because they lack conscious moral deliberation or cannot entertain moral propositions, we could run the risk of excluding much of human everyday moral interactions as well. Furthermore, even if typical human adults are capable of conscious deliberation, it is far from clear to what extent we rely on these capacities when engaging in morally relevant behavior (see also Paper II and 3.3). In this way, this debate connects to some of the issues made apparent in discussions about nonstandard human moral agency. Because of these considerations, I apply a practice-focused conception of moral agency to the case of canid social play (Paper II), and to some forms of human-animal multispecies social interactions (Paper III).

Similar to the artificial moral agency debate, the possibility of moral agency in nonhuman animals raises various additional practical questions. If some non-

[69] S. Fitzpatrick draws on Stich's (2009) and Nado and colleagues' (2009) distinction between *natural kind* and *conceptual analysis* approaches to define morality.

human animals are moral agents, should they be included in moral responsibility practices together with humans? Would it make sense, and be fair, to hold other animals morally responsible for their behavior? The possibility of multispecies moral responsibility practices is briefly discussed in the next section and is the main focus of Paper III.

Discussions about animal moral agency also highlight possible normative considerations in light of a possible link between moral agency and moral patiency (see this chapter, 4.3 and Chapter 5). In Paper IV, I claim that there seems to be a moral psychological link between moral agency and patiency in virtue of a distinct other-regarding perspective inherent to the participant stance. This link provides a normative reason to see all moral patients as partial moral agents, I argue.

4.3 Exempting Practices

This last section considers general justifications of, as well as worries about the objective stance. As mentioned, the objective stance is the perspective moral agents are assumed to take towards agents who are temporarily or permanently exempted from attributions of moral responsibility. Therefore, questions regarding exemptions primarily emerge in relation to nonstandard agents, such as small children, adult persons with allegedly moral agency undermining features or conditions, nonhuman animals, and artificial intelligence entities.

As we will see, our exempting practices are typically justified on rational and normative grounds. However, some authors reject the assumption that being ineligible for moral responsibility makes one an apt object of the objective stance. While I agree with many of the reasons provided in these discussions, I believe that an important piece is missing. I aim to bring this aspect to light by recapitulating the argument developed in Paper III and in doing so paving the way for Chapter 5.

4.3.1 Justifying the Objective Stance

Central to Strawson's account is the claim that we relate to other agents from two distinct perspectives or standpoints. The participant stance is the engaged, second-personal perspective we take up when we respond to the conduct of others with reactive attitudes. These attitudes are ultimately rooted in a demand for due regard or concern. The objective stance, on the other hand, is the perspective we adopt when we exempt someone from moral responsibility practices. To take this stance

is to see and relate to the agent and their conduct in terms of training, treatment, or manipulation.

The participant stance is therefore our default perspective to other moral agents. While we can sometimes adopt the objective stance for reasons beyond moral responsibility, such "as a refuge, say, from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity", as humans we cannot do so "for long, or altogether" (Strawson, 1962/1982, p. 67). However, with regard to humans and other beings who are not eligible for the reactive attitudes, the adoption of the objective stance is instead assumed to be justified.

An evident type of justification for our exempting practices concerns blameworthiness. A common assumption is that an agent is blameworthy for an action only if blaming responses, such as resentment or indignation, are appropriate to them for that action. The appropriateness of blaming responses has been suggested to be a question of backward-looking considerations, such as fairness (Wallace, 1994), accuracy or fittingness (Rosen, 2015; Shoemaker, 2015; McKenna, 2012), or felicitousness (namely, that the response is meaningful or intelligible) (Watson, 1987/2004; McKenna, 2012; Macnamara, 2015a). But the appropriateness of blame is also understood in terms of forward-looking considerations, such as its agency cultivating or scaffolding function (Vargas, 2013; McGeer, 2019).

Given these grounds of holding someone responsible, it would only be appropriate to blame someone who is eligible in the right way. For instance, fairness requires that the agent is capable of transgressing moral expectations, accuracy or fittingness (in terms of appraisals that track *slights*, blameworthy behavior, and so on.) require that the agent is capable of manifesting substandard regard, felicitousness requires that the agent is capable of uptake, and forward-looking considerations require that the agent is capable of adjusting, calibrating, or regulating in the requisite way. Conversely, the objective stance is appropriate or fitting to nonhuman animals, small children and the "mentally deranged" (Strawson, 1962/1982, p. 68), because they are assumed to be ineligible in one or several of the mentioned senses.

In this sense, taking the objective stance to someone is to view their behavior as, in some important sense, devoid of the kind of moral charge or meaning that we usually perceive in, and attribute to, the conduct of typical adult humans (see McKenna, 2012). Trees can fall on people and injure them, and cars can have motor failures and surely be a source of irritation. But we do not believe that it is reasonable or appropriate to see or treat trees or cars in any other way than as

natural elements, whose behavior can be explained and interpreted in mere causal terms.

Likewise, while small children and nonhuman animals may cause us annoyance or even harm, it just does not make sense to morally react to them. Strawson writes that "[i]f your attitude towards someone is wholly objective, then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with him" (Strawson, 1962/1982, p. 66). Resentment toward such beings, while sometimes understandable, does not have an apt recipient or target.

A further way to justify the objective stance, that often builds on the former, is to say that it would be unfair to view and treat some beings as appropriate targets of, say, blame. This is because "it does not seem fair to demand that people comply with such obligations unless they have the general ability to grasp those reasons and to regulate their behavior accordingly" (Wallace, 1994, p. 161). It would also be unfair to demand that they answer for their conduct if they are unable to provide such answers (Hutchison, 2018). In addition, some argue that it would be unfair to inflict "the emotional content of the demand—particularly the "sting" of resentment" on those who lack the capacity to understand moral demands (Hutchison, 2018, p. 218).

Adopting the objective stance is therefore thought to allow us to accommodate certain populations without imposing unreasonable burdens or unfair demands on them. The objective stance is therefore an attitude/stance which we *should* take toward some humans and nonhumans to whom reactive attitudes are not suitable or fair responses. It is a stance from which we avoid wronging or harming some moral patients (in part) *because* we refrain from treating them as targets or recipients of moral reactions or assessments. The objective stance is, in other words, assumed to be compatible with, and perhaps even necessary for, seeing and treating fairly persons and other beings who lack the required features of moral agency fairly.[70]

However, worries have been raised about the assumption that an objective stance to putatively moral agency undermined people or entities is fair and reasonable. For instance, Kennett argues that when we take a wholly, or even primarily objective stance toward mentally ill people, we deny them "psychological

[70] The next chapter (5) develops in more detail the mentioned implication that the objective stance is compatible with a type of moral consideration. See also Sommers (2007) for examples of views that claim that the objective stance is compatible with (and even necessary for increased) compassion.

visibility", and that doing so "further depletes or undermines their picture of themselves and their efficacy as agents." (Kennett, 2009, p. 111). Similarly, Pickard (2014, 2017) and Pickard and Ward (2013) question the assumption that holding responsible those who fail to meet traditional requirements of moral agency merely amounts to subjecting them to an unnecessary burden. They argue that "in holding service users with disorders of agency responsible, we treat them as one of us—as belonging with us, as equals." (Pickard & Ward, 2013, p. 1149).

How can one make sense of these concerns and claims given that the objective stance is also assumed to be reasonable and fair to certain populations? A possible route would be to take a closer look at the various attitudes and behaviors involved in the participant and objective stance. After all, the participant stance does not merely impose harms and burdens. It goes without saying that relating to others as moral agents also involves positive attitudes and responses, such as gratitude, admiration, esteem, and praise.

What is more, a common assumption is that holding responsible implicitly assumes, and conveys, a particular kind of respect or regard (see, for example, Jeppsson, in press). For example, McGeer states that "reactive attitudes communicate a positive message even in their most negative guise—even in the guise of anger, resentment, or indignation. The fact that we express them [moral responsibility reactions] says to the recipients that we see them as individuals who are capable of understanding and living up to the norms that make for moral community" (McGeer, 2012, p. 303). Similarly, Wolf (2011) thinks that:

> in revealing one's anger (or resentment or indignation) toward a person, one shows that one regards the person *as* a person, and as a member or potential member of one's community (at the relevant level of intimacy). Getting angry, as opposed to withdrawing one's trust, shows that one does not regard the person exclusively with the objective attitude[13](Wolf, 2011, p. 339).

The absence of these positive aspects of the participant stance becomes evident when we consider what exempting someone from moral responsibility practices conveys. For instance, A. M. Smith (2005) claims that taking the objective stance to someone suggests:

> that she is a passive victim of forces beyond her control, someone to be pitied and treated, perhaps, but not to be reasoned with or regarded as an appropriate participant in practices of interpersonal justification…. it should be clear that being denied responsibility for one's attitudes has its costs. Such denials can be deeply patronizing and disrespectful, and we should not be too

> eager to resort to them, either in our own case or in our treatment of others (A. M. Smith, 2005, p. 269).

In this way, taking the participant stance to someone involves a type of respect for or recognition of them qua morally responsible agent. One way of making sense of this type of respect seems to be found in what Darwall (1977) calls *appraisal respect* or what Hudson (1980) refers to as *evaluative respect*. Such respect involves appraising an agent for her moral conduct or character and is therefore something one earns or deserves in relation to meeting certain (moral) standards. But if it is the case that "the participant stance conveys respect for the moral capacities of those toward whom it is taken" (Hutchison, 2018, p. 214), it does not necessarily lend support to Kennett's and the others' worry.

While it may be the case that we exempt some populations from eligibility for moral appraisal, this may be reasonable and fair in virtue of relevant differences. Consider, for example, the expectations we may have toward an elite athlete as opposed to a non-athlete. The skills and abilities of the elite athlete seem to make it appropriate for us to expect a certain level or degree of athletic performance. However, these expectations constitute a double-edged sword: when our expectations are met or exceeded, we may of course express gratitude, praise, or even admiration. However, when the athlete performs below our expected standard (without there being any excusing conditions), we may instead react with disappointment and even anger.

Of course, one may still follow McGeer's and Wolf's contentions and say that even in their most negative manifestation, negative appraisals of athletic output tell the recipient that we view them as a highly competent or skilled athlete, as someone who is, generally speaking, eligible for such appraisals. Otherwise, we would not have bothered to subject them to such assessments in the first place. We thus have a certain evaluative or appraisal respect for them qua athlete. However, because the relevant features are lacking in the non-athlete, it would be unreasonable and unfair to subject them to the same expectations regarding athletic performance. Hence, while the non-athlete in some sense *misses out* on the positive reactions and attitudes of being regarded an elite athlete, this is because they lack the skill set or competence required for being subjected to the expectation or demand to begin with. They are therefore aptly disqualified.

Analyzing the potential benefits of being included as moral agent in terms of appraisal or evaluative respect would therefore support the claim that the objective stance is unproblematic in the case of appropriately exempted beings. They simply lack the features, properties, or competence required for being assessed and

engaged with in that particular way. A. M. Smith's account (2005) of what the objective stance suggests about its object, while true for moral agents, is then not applicable to beings who are aptly exempted. Taking the objective stance to someone who lacks moral capacities does not seem problematic, as there is nothing for which the exempting agent would fail to express respect or regard.

4.3.2 A Wider Appreciation of the Participant Stance

Nevertheless, there seems to be something to Kennett's and others' worry that cannot be entirely dismissed by appeal to appraisal or evaluative respect. For instance, Kennett (2009) argues that "[i]f reciprocal relations are, as Strawson suggests, the ordinary basis even for respect and goodwill, then we would expect to find, what we largely do find, that mentally ill persons are treated with less respect and less goodwill than other adult members of the community…" (2009, p. 105).

Shoemaker (2022) echoes this concern when he stresses the need to "distinguish between the interpersonal and accountability communities" (2022, p. 52). The former, Shoemaker argues, are not reducible to "the reactive attitudes responding to accountable agency." (2022, p. 52), but also involve things like love, compassion, and sympathy. Likewise, Wolf thinks that:

> we would do best to understand 'participation' broadly, as referring not necessarily to membership or even potential membership in … a community whose sole purposes are moral and political. We should rather understand the participant stance as involving the idea that the individual in question is 'one of us' in some wider or other sense. (Wolf, 2015, pp. 132-3).[71]

The worry, then, does not seem to simply be about missing out on being considered eligible for moral appraisal and evaluative respect. Rather, the concern seems to be that the participant stance involves attitudes and behaviors that go above and beyond reactive attitudes and appraisal respect. For example, exempting someone as moral agent may mean denying them good will, basic forms of respect, concern, compassion, or love that are not reducible to the attitudes and behaviors implicated in holding responsible. Likewise, some authors in the animal moral agency debate suggest that, while nonhuman animals may not be full-blown moral

[71] Note that these worries relate to, and imply, some of the claims found in discussions about the practical feasibility of adopting a general, or global, objective stance (Nelkin, 2011, Ch. 2; G. Strawson, 2010, Ch. 5; Watson 1987/2004, pp. 255-8; Sommers, 2007).

agents, there seem to be certain harms involved in denying them moral agency (or moral subjecthood) altogether.

For example, Rowlands (2012) argues that to treat a being with respect requires one to "treat it in such a way that it is able to exercise those capabilities, and so live a flourishing life." So, it "is important to them in the sense that it is *in their interests* for the others with whom they must deal to ask and answer this question [whether other animals are moral agents] correctly." (2012, p. 250). Similarly, Monsó and colleagues (2018) claim that, "[i]t seems, indeed, plausible to consider that the ways in which members of a species can be *harmed* make them *vulnerable* in certain specific ways, and, in turn, shape the kinds of *duties* we might hold towards them." (Monsó et al., 2018, p. 286; see also Benz-Schwarzburg & Wrage, 2023).

In light of these considerations, there appears to be certain aspects of, or attitudes pertaining to, the participant stance that are valuable, and perhaps available, irrespective of whether an entity is fully eligible for moral responsibility or not. Furthermore, this assumption finds support in the fact that a common type of proposal put forth by these authors is that we can see and treat responsibility exempted populations in ways that acknowledge and secure certain goods or benefits, while at the same time avoid imposing on them the unfair burden of, for example, blame or punishment.

For instance, Hutchison (2018) suggests that respect for a person's agential capacities can come in degrees, and as mentioned earlier, Pickard (2014, 2017) and Pickard and Ward (2013) recommend "responsibility without blame", that is, moral responsibility responses void of the affective and hurtful dimensions, as a means for clinicians to emancipate service-users with "disorders of agency" (Pickard & Ward, 2013, p. 1134). Similarly, Kennett claims that while mental illness should inhibit blame, it should not inhibit the participant stance. Given the importance of social relations for developing and supporting agency, "the default stance in both personal and professional dealings with those suffering a mental illness or disorder, as with anyone else, must be the participant stance." (Kennett, 2009, p. 112).

Considering animals, a common type of suggestion involves separating attributions of negative moral responsibility from praise (Borchert & Dewey, 2023). One reason is that this provides a way of recognizing the virtues of nonhuman animals, or to treat them with the respect they deserve (Wrage, 2022), without the harms or burdens involved in ascriptions of negative moral responsibility (Delon, in press). For instance, Burgis (2018) argues that animals are capable of "performing actions with positive moral worth" (2018, p. 130) despite not being

fit targets of ascriptions of negative moral responsibility. And Sapontzis (1980, 1987), drawing on a broadly Kantian notion of moral patiency, claims that because many other animals are able to act virtuously, they are "beings who must be respected" (1980, p. 51). Respecting animals in this way is argued to mean that they "must be treated as ends in themselves" (1980, p. 52).

Following a Scanlonian (1998), line of reasoning, Rowlands (2012) argues that the distinctive moral capacities of moral subjects demand "moral respect" (2012, p. 254) which goes above and beyond the considerations owed to mere moral patients. Monsó and colleagues (2018) develop this idea further and suggest that because some nonhuman animals are moral subjects, we may owe it to them that they can exercise their moral capabilities with conspecifics, for instance, by allowing social animals to engage in caring behaviors. The authors therefore criticize pure experiential approaches to animal welfare on the basis that such views fail to account for the kinds of harm imposed on nonhuman animal moral subjects in, for example, industrial animal agriculture (see also Wrage, 2022).

4.3.3 Concluding Remarks: the Case for Distinct Participatory Roles

While I am sympathetic to the mentioned arguments that question the appropriateness or justifiability of a wholly objective stance to nonstandard populations, I worry that many of them overlook a fundamental way in which one can include and exempt others. According to most of the arguments described, moral agency designates an entity's general eligibility for moral assessment and ascriptions of moral responsibility.

In the human-centered debate, being included as a moral agent is to be seen and treated as a potential target or recipient of moral reactions and assessments. In this way, the participant stance simply is the perspective from which we see others as such targets or recipients of moral appraisal.[72] The upshot is that the potential harms of the objective stance are taken to arise from the suspension of ascriptions of moral responsibility, like other-directed reactive attitudes.

While I follow the lead of the authors mentioned above, I hope to develop and improve the broad position they endorse. I argue that by focusing on the notion

[72] Likewise, authors in the animal-centered discussions, while more focused on the positive aspects of moral appraisal, limit their analysis to attributions of positive moral responsibility, moral virtues, or implications of moral capacities for flourishing. While there are some statements made about respect, these seem to be most closely linked to the type of appraisal respect mentioned earlier.

of moral agency as a type of eligibility for attributions or ascriptions of moral responsibility, one runs the risk of overlooking important attitudes and perspectives involved. These aspects are brought to light in Paper III, where I argue that seeing or exempting someone as moral agent often involves more than merely seeing or exempting them as morally responsible. Moral agency, conceived of as participation in communicative inter-relational practices, involves participating in (at least) two distinct roles or positions. These roles or positions, in turn, involve distinct participatory standpoints.

One can participate as a potential target or recipient of ascriptions of moral responsibility, that is, a *moral defendant*. But one may also participate as the source or addressor of moral claims and demands, that is, a *moral claimant*. Thus, the participant stance is not one, single, perspective. Rather the standpoint seems to be comprised of (at least) two distinct perspectives: one from which we see and treat someone as eligible for reactions and ascriptions of moral responsibility, and another from which we see and treat someone as a potential source or maker of assessments or ascriptions of moral responsibility.

Attending to these distinct participatory roles, and their respective aims, makes explicit that the participant and objective stances may involve including or exempting someone in two separate ways. For that reason, when assessing the permissibility of our exempting practices, we should not stop at identifying potential injuries, wrongs, or harms due to exempting from ascriptions of moral responsibility. This is an approach that rests solely on a *defendant* account of moral agency. As such, it represents a one-sided strategy that may provide limited, or even false, results about the kinds of harms of exemption and the extent of these harms.

Making accurate analysis of the adequacy and permissibility of an objective stance requires us to also consider the implications of exempting someone as moral claimant. Therefore, the next section picks up the defendant-claimant distinction developed in Paper III, and links this to claims made in Paper IV via a discussion about the possible connections between moral patiency and moral agency. In particular, I consider how seeing and exempting as moral claimant appears to involve two distinct conceptions of moral patiency.

5 Moral Agency and Moral Patiency

The aim of this chapter is to provide more extensive background to, and further elaborations on, the arguments and claims developed in Paper III and Paper IV. As mentioned at the end of the previous section, I believe that discussions about the justifiability and potential harms of exempting from moral agency have overlooked important aspects of moral participation. Recognizing the claimant dimension of moral agency, and how it connects to moral patiency, I argue, can make explicit important implications of setting standards for moral agency (Paper IV). Attending to these implications, in turn, allows us a more thorough understanding of the nature and scope of the benefits and harms associated with including or exempting someone as a participant in moral responsibility practices (Paper IV).

This chapter begins by considering the concept of moral patiency and its various suggested grounds. The subsequent section accounts for the ways in which moral agency and moral patiency are commonly assumed to be related. This section also introduces the traditional Kantian account where moral agency and patiency are assumed to be two sides of the same coin. The third section goes on to examine some contemporary Kantian and broadly Kantian accounts of two distinct grounds for, or types of, moral patiency designed to avoid forceful objections to traditional Kantianism. These accounts help to elucidate arguments for and worries about the objective stance found in the practice-focused literature on moral responsibility and agency as a standard way of viewing beings that are exempted from moral agency. In the following section, I then go on to identify and challenge three distinct arguments in favor of a purely objective stance to moral agency exempted moral patients. I argue that since the participant and objective stances assume distinct conceptions of moral patiency, and since these conceptions dispose the stance-taker in morally relevant ways, the arguments developed in Papers III and IV gain further support.

5.1 Moral Patiency

This section considers the basic philosophical concept of moral patiency, its suggested grounds, and the particular account of moral patiency assumed in this thesis.

5.1.1 The Concept of Moral Patiency

Moral patiency and moral agency are both central but (it is usually assumed) distinct philosophical concepts. Moral patiency denotes an entity's status or standing as morally significant in its own right. This is usually taken to mean that the being in question has interests that ground moral reasons or requirements for how others are to treat that being.[73] As mentioned in the outset of Chapter 2, moral agency denotes a being's eligibility for ascriptions, assessments, and responses of moral responsibility. That is, moral agents are beings who can do right or wrong and who can be held morally responsible for their conduct. By contrast, moral patients can be targets of right- or wrongful conduct at the hand of moral agents (Jaworska & Tannenbaum, 2023; Gruen, 2021; Warren, 1997).

5.1.2 The Grounds of Moral Patiency

Like with moral agency, views on the grounds and requirements of moral patiency differ. For instance, some accounts posit that for an entity to be a moral patient, it has to be a member of a certain group, for example the biological category designating the human species (Dworkin, 1993; Benn, 1967). Other theories claim that moral patiency requires having certain cognitive (intellectual and emotional) capacities or features. A widely held view of the latter kind is that sentience or phenomenal consciousness is necessary and sufficient for (at least basic) moral patiency (Varner, 2001).

A popular cognitive capacities view is found in accounts that argue that, while sentience is sufficient for a basic form of moral patiency, more advanced cognitive capacities may grant a being higher or even *full* moral patiency (in terms of the force and type of the moral requirements that the patiency implies for moral agents). For instance, according to McMahan's (2002) two-tiered account, sentience is sufficient to grant an entity some protection but allows that their interests be considered from a purely consequentialist perspective. As such, merely

[73] What I refer to here as moral patiency is sometimes also referred to in terms of moral status (Jaworska & Tannenbaum, 2023; Warren, 1997) or moral considerability (Gruen, 2021).

sentient beings can be killed for the greater good. However, having advanced cognitive capacities, like self-awareness, grants a being respect, which means that they are owed to be treated as *inviolable* (McMahan, 2002).[74]

According to yet another cognitive capacities view, even basic moral patiency is thought to require sophisticated cognitive capacities. The most famous account of this type is the one proposed by Immanuel Kant (1785/1996, 1788/1996, 1797/1996. In short, Kant argued that moral patiency requires the capacity for autonomy and the capacity to set ends (Korsgaard, 1996; Hill, 1997; Wood, 1999). An important feature of the orthodox Kantian position is that both of the mentioned capacities are assumed to be required for moral agency. Hence, moral agency is assumed to be a prerequisite for any type of moral patiency. While I believe that the Kantian account is flawed (and will state my reasons later), I also believe it can serve to shed light on potential harms of the objective stance. I will therefore return to Kantian accounts of moral patiency and moral agency below (5.2.2).

5.1.3 Sentience, Interests and Obligations

In this thesis, I will assume that sentience is necessary and sufficient for (at least basic) moral patiency. To say that a being is sentient can, on a broad definition, be taken to mean that the being has phenomenal consciousness. This means that (some of) the entity's psychological states have the property of which there is "something that it is like" to be in those states (Nagel, 1974, p. 436) or, to put it differently, that the mental state in question has a subjective, or experiential, quality (Block, 1995; Van Gulick, 2022; Tye, 2021).

Some familiar examples of phenomenal mental states are tasting umami, hearing thunder, seeing purple, feeling nostalgic, and experiencing sadness. Other, probably less familiar, examples of possible phenomenal mental states can be found by considering the *Umwelt* (von Uexküll, 1934/2010) of other animals. For instance, some nonhuman animals appear to sense the Earth's magnetic field (Mouritsen and Ritz, 2005; Johnsen & Lohmann, 2005), see heat (A. L. Campbell et al., 2002), smell the passage of time (Horowitz, 2016), hear the shape of their surroundings (Jensen et al., 2005), or feel the electrical currents emitted by other living beings (Sisneros & Tricas, 2002).

---

[74] Similar claims about distinct kinds, or degrees, of moral patiency or the legitimacy of differential treatment have been made in relation to, among other things, the capacity to form future-oriented plans and desires (Singer, 1993) and having foresight (Rachels, 1990; Regan, 1983/2004; DeGrazia, 1996, 2008).

However, the kind of sentience typically assumed to be of ethical significance refers to the capacity of experiencing phenomenal mental states that are negatively or positively valenced. In other words, the capacity to have subjective experiences that feel good or that feel bad, such as anxiety, pain, boredom, comfort, happiness, pleasure, and so on (DeGrazia, 1996; Duncan, 2006; R. C. Jones, 2013; Browning & Birch, 2022). A prominent defender of the sentientist view of moral patiency was the eighteenth-century philosopher Jeremy Bentham, who famously wrote regarding the question of moral consideration of, for instance, nonhuman animals: "the question is not, Can they *reason*? nor, Can they *talk*? but, Can they *suffer*?" (Bentham, 1789/2017, Ch. 17, note 1).

On this view, sentient creatures are assumed to have an interest in avoiding certain states (such as pain) as well as an interest in the promotion of other states (such as pleasure). Consequently, such beings can be benefitted or harmed by experiencing "positive or negative impacts on their interests" (Gruen, 2017, p. 91). This provides all moral agents with non-instrumental reasons to morally attend to their treatment of sentient beings. Contemporary proponents of sentientism are found among utilitarian (Singer, 1975, 1993) as well as Kantian inspired philosophers (Regan, 1983/2004; Korsgaard, 1996, 2018a).

While there is scientific consensus on the sentience of mammals, birds, and even some mollusks, like octopuses (Low et al., 2012) there is still controversy regarding whether, for example, insects and crustaceans (see Browning & Birch, 2022), not to mention, whether artificial entities can experience phenomenally conscious states (Dehaene et al., 2017). Although I will not discuss these controversies much further, it is worth reiterating the idea of *multiple realizability* as it bears on the question of whether sentience can be attributed to entities with physical properties that diverge from the mammalian, vertebrate, or even biological norm.

According to the idea of multiple realizability, a psychological state, like subjective pain, can be realized by distinct physical kinds (Putnamn, 1967), such as neurons, electronics, or green slime (see Bickle, 2020). Multiple realizability is already assumed in attributions of sentience to organisms, like octopuses, who have relatively advanced cognitive abilities and complex behaviors, but very different neurological underpinnings from vertebrates (see, for example, Birch et al., 2021). Hence, assuming multiple realizability, one cannot deny the possibility of sentience in an entity merely on the basis of its physical constitution. This, of course, has implications for the question of artificial moral patiency (see, for

example, Danaher, 2020) and connects to the pragmatic epistemic view discussed in the previous chapter (4) and in Paper I.

5.2 Moral Agency and Moral Patiency

So far, the distinction between moral agency and moral patiency has been characterized as fairly straightforward. These two concepts are generally assumed to refer to distinct types of morally relevant group memberships: one regarding the group of beings who can act morally right or wrong, and the other regarding the group of beings who can be treated rightly or wrongly. As we will see, however, there are various views and related complications about how these memberships are related, both conceptually and in practice. While some theories retain the commonly assumed distinction between moral patiency and agency, others do not.

5.2.1 Moral Agency and Sentience

Although not always explicitly stated, moral agents are often assumed to also meet the requirements of moral patiency in practice. While moral agency and moral patiency may be distinct concepts, they appear to overlap to a significant extent in real life cases. This relationship, while not necessarily given, appears to be true if we look at paradigm examples of moral agents. The standard moral agent in philosophical literature is assumed to be a typical adult human. And humans, in general, certainly have morally significant interests in virtue of, among other things, being sentient. That is, according to one type of view, moral patiency is a status that happens to be possessed by all creatures who are also moral agents.

According to other accounts, however, the link is tighter. The features or abilities underlying moral patiency are then argued to be the same as some of the features or abilities required for moral agency. Moral agency is thus thought to *necessarily imply* moral patiency in virtue of some other feature or property that, in turn, also grounds moral agency. For example, some so-called sentimentalist accounts of moral judgment and moral motivation argue that certain phenomenal states are central to moral thinking and behavior. For instance, moral motivation is argued to require empathic identification or feelings of sympathy, implying sentience (Kauppinen, 2022). In effect, the capacity to experience certain phenomenal states that will confer moral patiency is assumed to be necessary for moral agency.

As mentioned in Chapter 3, various accounts of moral agency similarly assume that phenomenal states play a central role in our practices of holding each other

morally responsible. Blaming responses, for instance, are assumed to involve reactive attitudes like resentment or guilt, which are often characterized as, in part, comprising of phenomenal states (see Paper III). On such theories, participating in moral responsibility practices thus requires that one is prone to respond with such phenomenally conscious attitudes to perceived wrongdoing. Hence, assuming that sentience is a sufficient condition for moral patiency, the mentioned accounts of moral agency will have the implication that all moral agents necessarily meet this condition.

The assumption that moral agency presupposes consciousness, and thus the features or capacities that ground moral patiency, can be found in some standard accounts of moral agency discussed in Paper I. For example, some believe that the traditional control and knowledge conditions for moral agency both presuppose consciousness. For an agent to have control, they need to be able to make conscious rational decisions, as opposed to merely behave rationally, or deciding in a way that do not presuppose subjectively experienced reasoning. For an agent to be able to navigate the moral landscape in the first place, they need to be able to consciously grasp and apply moral concepts to have the required knowledge (Himma, 2009).

The significance of phenomenal consciousness for moral agency, while widely assumed, is likewise subject to great dispute in the artificial moral agency debate analyzed in Paper I. There, Munthe and I concur with the arguments put forth by those denying phenomenal consciousness as a distinct requirement for moral agency. Assuming that we want to maintain our current pragmatic practices of ascribing moral agency and responsibility, there are pragmatic-epistemic reasons to abandon the consciousness requirement and instead focus on observable features that may depend on, or indicate, consciousness and that are relevant for participation in moral responsibility practices. Such measurable features could, for example, be states or behaviors indicative or constitutive of resentment, guilt, sympathy, concern, and similar attitudes (see also Paper I, sec. 3).

Before concluding this section, it is worth noting that even if all moral agents necessarily have the features required to qualify as moral patients, not all moral patients necessarily qualify as moral agents because the conditions that are sufficient for moral patiency are unlikely to suffice for moral agency.

5.2.2 Mere Moral Patients

Given that the capacity for phenomenal consciousness is the basis of moral patiency, and assuming that moral agency requires features or capacities above and beyond sentience, there are many beings in the world who seem to be *mere moral patients*: beings who have morally significant interests but lack the features or abilities required for membership in the moral agency club. For example, cats and infants are moral patients in virtue of meeting the requirement of sentience. However, they do not seem to be appropriate targets of, say, blame. Thus, a common assumption is that moral patients are owed certain treatment from moral agents, without themselves necessarily being subject to any such requirements (Gruen, 2021). After all, the notion of moral patiency is what renders fairness- or cruelty-based arguments for exempting certain moral patients from moral responsibility practices relevant. Whether some particular treatment of an entity is permissible or justified depends, in the first instance, on whether the entity in question is a moral patient (see previous chapter, 4.3).

However, while the importance of the notion of mere moral patiency is widely embraced by people in the contemporary moral patiency and agency debates, I wish to call it into question, and I have started to argue to this effect in Papers III and IV. Before elaborating further on this challenge, I will, however, first take a detour via broadly Kantian accounts of moral patiency as these likewise deny the existence of *mere moral patients,* but for very different reasons from mine. In addition, contemplating these accounts, I will argue, helps bring to light premises and assumption implicit in justifications of, as well as worries voiced against, the objective stance (see previous chapter, 4.3). These aspects will be of importance when I develop my own fleshed out argument for why all moral patients should be granted some moral agency (see 5.4 below).

5.2.3 Kantian Views on Moral Patiency

According to Immanuel Kant, a person is "a being altogether different in rank and dignity from *things*, such as irrational animals, with which one may deal and dispose at one's discretion" (1798/2010, p. 239). This is because, on Kant's view, experience-based attributes, such as sentience, are not the basis of personhood to begin with. Instead, for someone to be a person and therefore matter morally in their own right they must, in addition to an *animal nature*, also have a *rational nature*, which involves the capacity to will freely from reasons, set ends and assess their means, and regulate their behavior according to principles. On Kant's own

account, to be a moral patient requires all of these specific features, features which he also tends to assume set humans apart from other animals.

In this way, it is not just that moral patiency and moral agency may overlap in practice, or that moral agency requires features involved in moral patiency. Instead, Kant's view makes the stronger claim that moral patiency, in terms of mattering morally in one's own right, and moral agency, are two sides of the same coin. Moral agency, understood as the capacities to legislate oneself and conform to moral principles (Korsgaard 1996; Hill, 1997; Wood, 1999), is constitutive of moral patiency by granting persons the dignity reserved for moral agents and thus making them deserve respect (Wasserman et al., 2017).

A possible implication of the Kantian view is that artificial entities could have, say, rights, if they were moral agents. Depending on the requirements of moral agency, this might therefore present a potential route to moral patiency for artificial entities that circumvents the requirement of phenomenal consciousness or sentience. However, another implication, and the one I will focus on here, is that there are no moral patients who are not moral agents. The category of *mere moral patients* is empty. For instance, because cats, infants, and moral agency exempted adults lack moral agency, they are not moral patients either (Kain, 2009).

However, this exclusion of moral agency-exempted beings from moral patiency does not necessarily mean that moral agents may treat cats, infants, and moral agency exempted adults however they want. Kant famously argued that beings who are not moral patients can still be morally considerable in an indirect sense. While such beings are, in fact, mere things and cannot be owed anything directly, Kant believed that one should avoid treating them cruelly as this may "dull" our human feelings and cause us to treat other moral agents badly (Kant, 1797/1996, p. 564). We should therefore "practice kindness towards animals, for he who is cruel to animals becomes hard also in his dealings with men" (Kant, 1780-1/1963, p. 240).[75]

Our duties or obligations to moral agency exempted creatures and humans may also be derivative in the sense that we owe it to other moral agents who value such beings to treat the former decently. This idea is described by Carruthers (1992), who states that because animals are not moral contractors, they, "like buildings,

[75] O'Neill (1998) argues that, despite appearances, our indirect duties to animals may in practice imply the same type of considerations as welfare-based accounts, which "is not a trivial protection" (1998, p. 223). She writes: "in allowing that harming non-human animals is an *indirect* violation of duties to humanity Kant endorses more or less the range of ethical concern for non-human animals that more traditional utilitarians allowed: welfare but not rights." (O'Neill, 1998, p. 223; see also Denis, 2000).

would have no direct rights or moral standing. Rather, causing suffering to an animal would violate the right of animal lovers to have their concerns respected and taken seriously." (1992, pp. 106-7). In a similar vein, Warren claims that infanticide is unacceptable only insofar as "there are other people who would … be deprived of a great deal of pleasure by its destruction" (1995, p. 453).[76]

Kantian instrumentalist considerations have also been raised in relation to artificial entities. Some worry that the way we treat and behave to machines and other artificial entities may impact the way we see and relate to other humans. In particular, that our treatment of sentient-like or human-like artificial entities as mere things or objects runs the risk of negatively influencing our treatment of humans. Conversely, seeing and treating, say, robots as moral patients may serve to protect against desensitization and therefore ensure morally acceptable behavior toward humans and nonhuman animals Hence, granting rights or protection to artificial entities may act as an instrumentally motivated safeguard against the mentioned risks (Darling, 2016).

Others, however, take the opposite position and argue that ascribing moral patiency to artificial entities that may appear to have the features required for moral patiency is the wrong type of solution to the mentioned worry. Assuming that perceiving and engaging in apparent harmful behavior toward such entities may desensitize humans, we should refrain from designing realistic humanoid robots that may *appear* sentient in the first place (Johnson & Verdicchio, 2018).

Of course, many, (myself included) find the traditional Kantian position on indirect duties to moral agency exempted sentient beings far from satisfying. In addition, many present-day Kantian philosophers appear to share this assessment. Some have even called it a "repugnant moral doctrine" (Hill, 1997, p. 58). It seems intuitively obvious that the reason why moral agents need to consider the interests of infants, cats, and moral agency exempted adults is, at least in part, that they are beings who are morally considerable in their own right. If we lived in a world where no moral agents cared for moral agency exempted beings, or where cruelty to such beings had no negative effects on our attitudes and behaviors toward moral agents (Wood, 1998; Skidmore, 2001; Calhoun, 2015),[77] there would, on Kant's view, be no duty not to, say, torture animals. Hence, the reason why we should refrain from harming moral agency exempted sentient beings surely cannot simply be a matter

---

[76] Similar indirect or instrumental reasons also figure in multi-level cognitive capacities accounts discussed earlier, such as McMahan's (2002).

[77] See Wood (1998) for a critical discussion of the moral psychological assumption inherent in Kant's argument.

of derivative or contingent instrumental reasons (Korsgaard, 2018a, 1996; Nussbaum, 2005, 2006; Wood, 1998; Talbert, 2006; Kittay, 1999).

Now, while I do oppose Kant's claim that moral agency exempted beings lack moral patiency, I still believe that the basic idea of duties to others as conditional on them having certain moral agency-related features, reflects important insights about human moral psychology and behavior: A typical way in which we come to perceive and respond to moral considerations is to see others as active and present agents, particularly as agents, in particular, as agents who can morally appraise us. That is, a central dimension of moral sensitivity and responsiveness involves perceiving and experiencing moral constraints as somehow originating from actual or prospective second-personal claims, demands, commands, objections, protests, et cetera.[78]

These moral psychological assumptions are central to the normative claim developed in Paper IV. As I hope to show, they likewise figure more or less implicitly in a variety of different contemporary theories of moral responsibility and agency. In particular, the mentioned moral psychological ideas play a fundamental role in philosophical social contract accounts inspired by Kant's moral philosophy and can thus be brought to light by considering these accounts.

5.3 Justification, Moral Sensitivity and Motivation

A common feature of many social contract accounts is the idea that reciprocal or mutual accountability relations with others is central to what we owe to each other and why we owe these things. This general idea takes two primary forms, contractarianism and contractualism. In the following I will briefly describe the former to put it to a side, and thereafter focus on the latter as a type of Kantian social contract theory which reveals interesting features of how and why we ascribe moral agency and patiency.

According to the Hobbesian (Hobbes, 1651/1997) line of social contract thought (also known as *contractarianism*), people are first and foremost self-interested and not intrinsically motivated by moral concern for others. However, given conditions of competition or scarcity, rational and self-interested agents are thought to understand that they each will benefit by agreeing to cooperate with

[78] This image of the reality of human moral psychology also fits the generally accepted notion of reciprocal altruism in evolutionary (moral) psychology (see Trivers, 1971 and Brosnan and de Waal, 2002).

others (Cudd & Eftekhari, 2021; Gauthier, 1986; Buchanan, 1975/2000). In this way, contractarianism holds that people are motivated to follow moral norms in virtue of a rational assessment of how to best reach their own goals, and that this assessment favors a strategy of cooperation and justice. Hence, sensitivity and responsiveness to moral considerations, including seeing and respecting others as moral patients, is claimed to be grounded in, and motivated by, purely instrumental reasons. The implication of this is that moral patiency presupposes rationality and self-interest, rather than sensitivity and responsiveness to moral considerations. Moral agency, therefore, is not really part of this picture at all albeit it may be a contingent secondary upshot of a social cooperation motivated by this type of purely instrumental reasons (as it may require accountability mechanisms to be managed). Since I am here interested in possible links between *moral* agency and moral patiency I will therefore set the contractarian view aside.

5.3.1 Contractualism

In contrast, contractualism applies a basically Kantian notion of the link between moral agency and patiency, as it grounds the latter in the former in virtue of a hypothetical agreement between all moral agents. Contrary to the contractarian supposition, contractualist accounts start from the assumption that people can be motivated both by self-interest *and* by respect for others. This is because moral agency involves the intrinsic desire to take into appropriate consideration the evaluative viewpoint of others. A rational moral agent will recognize that to have one's claim seen as a justified moral requirement, it must be weighed and assessed in relation to the actual or prospective claims of others. As such, "[it] follows that whenever you claim a right, you commit yourself to respecting the rights of others" (Korsgaard, 2018b, p. 20).

Hence, to respect someone means acknowledging them as authorized to run their own lives and mutually accountable for how they choose to live (Darwall, 2006), and thus view them as a source of actual or hypothetical (reasonable, legitimate) claims, demands, or objections that impose moral constraints (Scanlon, 1998, 2008; Darwall, 2006; Korsgaard, 2018b). In this way, the contractualist notion of moral patiency presupposes moral agency, according to contractualism, because a being's standing or authority as claim-maker (that is, someone who is a source of moral constraints on the conduct of others) presupposes the features required for standing in reciprocal relations of accountability. Only moral agents have the features or capacities required for imposing constraints on the actions of

others, since only such agents can recognize and be moved by reasons for or against actions in terms of, say, reasonableness (Scanlon, 1998, 2008). In this way, our obligations to others "are grounded in relations of reciprocal legislation" (Korsgaard, 2018a, p. 147).

The fundamental idea in all these accounts is that only beings who have what it takes to evaluate others, morally appraise their actions and holding them responsible, can impose moral constraints on others. That is, only moral agents have what is required for membership, or participation, in the moral community, agreement or relationship of mutual consideration. Only moral agents are sources of what in the literature is variably termed directed, authoritative, relational, agent-relative, or justice-entailing obligations, duties or considerations.

However, the assumption that in order to impose such constraints on others:

> a being must be able to have moral claims made on it (and hence be capable of moral responsibility) … would exclude any human being lacking the capacity to have moral claims made on them—not only individuals with radical cognitive impairments, but infants and young children as well. (Wasserman et al., 2017, sec. 2.1).

In addition, grounding rights in relations of mutual accountability means that "[r]ights protect a kind of liberty that the other animals do not and could not possibly have" (Korsgaard, 2018b, p. 11). Therefore, the mentioned accounts seem to encounter a problem similar to the one in Kant's original view: moral agency exempted beings are counterintuitively excluded from moral patiency.

5.3.2 Two Distinct Grounds of Moral Patiency

Perhaps for this very reason, the present-day variants of the contractualist idea tend to diverge from Kant's original view with regard to one important point: they leave room for other types of (non-instrumental) moral considerations beyond those grounded in moral agency. While only moral agents can have rights, be subjects of justice, be entitled to respect, be owed directed duties, et cetera, there are other types of moral considerations available in the case of moral agency exempted beings. In this way, these contemporary members of the Kantian and contractualist family hope to be able to account for our intuitions and views about the moral patiency of infants, cats and moral agency exempted adults, while maintaining that moral agency grounds a particular type of moral patiency that goes beyond the consideration owed to mere sentient beings.

The primary type of solution found in these accounts, is to distinguish between two grounds of moral patiency. On this view, moral agents are entitled to dignity, are right-bearers, and are owed directed duties of justice and respect. To respect someone implies that one recognizes a distinct form of authority, power, or standing that dictates our conduct toward and relations with others (see, for example, Darwall, 2006). However, despite the fact that beings like infants and nonhuman animals are claimed to be unable to reason morally or be part of reciprocal moral relations, they are thought to be eligible for moral consideration on (partially) different grounds. In this way, then, contemporary Kantian-inspired contractualists are increasingly recognizing "two distinct moral phenomena" (Alm, 2023, p. 16), where one is exclusively attributed to moral agents or persons and the other to beings who are exempted from agency and personhood.

For instance, some argue that because sentience implies that one has *a good*, merely sentient creatures may be owed direct duties of benevolence, which are grounded in a being's animal nature as opposed to her rational and moral nature (Korsgaard, 2018a). Others claim that moral agency exempted beings may be morally accounted for by being objects of concern or sympathy in virtue of having a welfare (Darwall, 2006). Hence, acting out of benevolence or sympathy for a being's animal nature or her welfare represents an additional way of valuing someone for their own sake. This type of non-instrumental moral consideration is assumed to be available in the case of moral agency exempted moral patients and does not involve seeing or treating them as a co-participant in moral responsibility practices.[79]

Another suggestion for how one could account for moral agency exempted beings or entities is that they could be included via a system of "trustees" (Scanlon, 1998, p. 183; Darwall, 2006, p. 29). For instance, Scanlon (1998) argues that, while it does not make sense to (actually or imaginatively) justify oneself to beings who cannot understand such justifications, we could perhaps indirectly ground the moral patiency of such beings in a contractualist framework, by justifying our treatment *about* such beings *to* other rational agents acting as *trustees* for the interests of the exempted beings. What is important, Scanlon believes that the considerations

[79] Various authors propose that Kant's moral philosophy seems to imply that we have duties *regarding* or *about* nonhuman animals despite not having any duties *to* them (see, for example, Denis, 2000). The to/about distinction, I believe, goes to the heart of the two different grounds of moral patiency found in many contractualist theories.

that such trustees could raise are "limited to objections based on experiential harms such as pain and distress" (1998, p. 184).[80]

Consequently, on contemporary Kantian contractualist conceptions of moral patiency, moral agency exempted beings cannot, in and by themselves be sources or originators of directed duties. They can however be considered as passive recipients of moral consideration, such as by being objects of concern, possibly via a *trustee* among moral agents. Adopting the terminology from Paper IV, I will henceforth refer to the two conceptions of moral patiency proposed by Kantian and contractualist theories as referring to the roles of moral *claimancy* and *wardship*, respectively. The *claimancy* role applies to parties to the moral contract or agreement, while the *wardship* role applies to moral patients who cannot be contractors. The objects to which these concepts refer to are thus moral *claimants* and moral *wards* (Paper IV).[81]

5.3.3 Claimants and Wards

Moral claimants are active, present participants or members of moral responsibility practices. In this way, seeing others as moral claimants implies adopting a particular type of the participant stance to them, or more precisely, what I in Paper III refer to as the *claimant-directed* participant stance. Moral claimants are seen as originators or addressors of moral claims and demands, and as having a certain standing or authority which makes their evaluative perspective (in terms of their actual or hypothetical objections, wants, goals, plans, opinions, preferences, et cetera) act as potential valid constraints on the conduct of others. In consequence, moral agents are required to consider the (actual or hypothetical) claims, demands, ends, objections, et cetera of moral claimants. In other words, moral agents are necessarily answerable and accountable to moral claimants.

Moral wards, on the other hand, are situated outside of the boundaries of participation in moral responsibility practices. They are absent from actual, prospective, or imagined moral communicative exchanges. Seeing others as moral wards therefore means exempting them as claimant participants in moral

---

[80] It is worth noting that Scanlon (1998) likewise raises the possibility that the scope of contractualism may not encompass all of morality. Hence, he suggests that there might be other kinds of obligations regarding nonhuman animals and other entities, who are not parties to the contract.

[81] I am here building on the terminology of Kymlicka and Donaldson (2017). Note, however, that they use *ward* and *wardship* as a second-class citizenship status rather than a type of moral patiency.

responsibility practices. Wards do not, and should not, figure as sources or addressors of objections, claims, demands, etc. in the practical deliberation of moral agents. However, wardship is still a membership in virtue of which the being in question has morally significant interests in terms of, for example, having a welfare, *a good* based on being an experiencing subject, et cetera. Such interests do pose constraints on the actions of moral agents, but the latter are not accountable and answerable to wards in any genuine sense. For example, it is generally assumed that one has duties in relation to pure wards, such as pets or infants, in the sense of being required to care non-derivately for their health and wellbeing.

The distinction between moral claimants and wards is essentially a question of differentiating between those moral patients who have the capacity or authority to make demands, impose corresponding directed-duties, and those who do not. The assumption that moral agency entitles a being to a particular type of moral consideration is, I believe, also central (but often only implicitly) to practice-focused accounts of moral agency and responsibility. For instance, Tognazzini states that "reciprocity, or at least the possibility of it, does seem to go to the heart of what Strawson calls the participant stance" (2015, p. 34). And Shoemaker connects responsibility to regard via reciprocity, by stating that "we can best capture our understanding of regard by restricting its scope to other agents, where these may be humans or (some) nonhuman animals. Indeed, this makes sense insofar as we think of accountable agents as accountable to others." (Shoemaker, 2015, p. 93). Lastly, Strawson writes that "The objective attitude … cannot include … the sort of love that two adults can sometimes be said to feel reciprocally for each other" (Strawson, 1962/1982, p. 66).

Recognizing the connection between the participant and objective stances and the contractualist suggestion about two distinct grounds of moral patiency can help elucidate some of the worries raised about the objective stance in the previous chapter (4.3). Opponents of the objective stance are right in pointing out that the implications and consequences of exemptions go beyond withholding reactions and ascriptions of moral responsibility. Including or exempting others as participants in moral responsibility practices likewise involves including or exempting them as moral claimants (Paper III), and therefore attributing one of two distinct types of moral patiency (Paper IV).

5.4 Questioning Pure Wardship

My view, as expressed in Papers III and IV, challenges the assumption that a pure ward perspective towards agents exempted from moral responsibility practices is justified. In this section I develop some of these arguments and relate them to the contractualist distinction between two distinct grounds of moral patiency. In particular, I will spell out and respond to what I take to be three related but distinct arguments in support of the case for pure wardship, which are extrapolated from both the contractualist and practice-focused literature.

5.4.1 The Argument From Symmetry

First, there is the argument from *symmetry,* which is described in more detail and responded to in Paper III. According to this argument, a pure ward perspective is justified towards moral agency exempted moral patients since they lack the internal features or capacities required for moral agency. The idea is that if a being lacks the features required for being held morally responsible, they will *necessarily* lack the features or capacities required to hold others morally responsible. Hence, this argument appeals to an assumption about how the features or properties that underpin moral agency as moral defendant are the same as (or overlap with) those that underpin moral agency as moral claimant (Scanlon, 1998; Darwall, 2006; Russel, 2004; McKenna, 2012).

In short, Paper III questions this argument by pointing out that, while the psychological underpinnings of the claimant and defendant dimensions are the same, an agent may not be symmetrically *eligible* as claimant and defendant in every given social context. One's eligibility for moral address and moral response is contingent on various factors pertaining to the more precise nature of the interaction. Therefore, asymmetry can show up *between* particular agents in particular circumstances.

I argue that dogs and toddlers, for instance, are asymmetrical, *claimant-heavy*, participants *in relation to* typical adult humans. The support for the prevalence of *asymmetrical responsibility relations*, I argue, shows why the argument from symmetry does not support the case for denying moral claimancy to agents who may not be capable of reciprocating moral claims and demands.

5.4.2 The Argument From Adequacy

Secondly, there is the argument from *adequacy*, according to which the welfare-based consideration available from within the ward perspective "fits the needs and interests" of exempted populations (Kymlicka & Donaldson, 2017, p. 842). In other words, moral agents are thought to be able to know, and direct themselves by, what they owe to moral agency exempted moral patients by the means available from outside of actual or prospective moral exchanges, namely, by seeing and relating to them as the proper objects of moral concern that they are. This assumed adequacy of a wardship stance is, I believe, a central, albeit seldom clearly disclosed, premise in justifications of the objective stance in practice-focused approaches to moral agency and responsibility (see the previous chapter, 4.3.1 and Paper IV). But, as we saw in the previous section, this assumption likewise figures in present day Kantian contractualist accounts of moral patiency.

In Paper IV, I challenge the argument from adequacy by arguing that a mere wardship stance to moral agency exempted sentient beings involves potential harms due to the implication of this stance to exempt them as moral claimants. This is because a pure ward perspective toward a moral patient, disposes the moral agent to overlook or disregard the actual, possible, and hypothetical claims, demands, objections, and so on, of that being.

Of course, assuming the trustee model may widen the range of possible moral considerations available to us. For example, this model might be taken to mean that we govern our behavior in light of claims, demands or objections that a representative actually makes, or could make, on behalf of some moral patient. Hence, the actual or imagined moral address of trustees may very well serve to scaffold moral sensitivity and responsiveness in relation to wards.

However, considering the reasons put forth by trustees, or the reasons one imagines trustees to put forth, cannot substitute the much more solid scaffolding potential of adopting a claimant perspective. When an agent sees a moral patient as a moral claimant, as opposed to a voiceless and passive ward, they see things, including their own conduct and attitudes, from a different perceptual stance. If the trustee model can be said to appoint some members of one's already recognized moral audience to represent the interests of absent moral patients, the claimant perspective grants entrance for those moral patients into the full realm of moral consideration. As such, this perceptive adds *new* members to one's moral audience.

A practical implication of this difference between the trustee model for pure wardship and a bona fide claimant stance is that the claimant perspective, but not the trustee model, can serve to make agents better positioned to perceive some of the behaviors of typically claimant exempt beings, such as very young children or nonhuman animals, as possible instances of moral address. As such, a pure ward perspective, even one supplemented with actual or imaginative address from an already recognized representative trustee, cannot make up for the range of considerations made visible by considering the evaluative perspective of the particular entity in question.

Hence, although animals and small children may not be capable of, or as proficient in, expressing claims or demands linguistically or of considering how rejecting or endorsing an action may impact others, this does not take away the fact that all sentient beings possess a morally significant evaluative perspective (Talbert, 2006). As such, they have interests also in the sense of subjectively aiming for, avoiding, endorsing, rejecting, preferring, planning for, or liking, some state over another. Hence, we have reason to consider their evaluative point of view when considering how to act, or we will risk culpable recklessness or ignorance (see Papers I and II for discussions regarding precautionary reasons to ascribe moral agency).

5.4.3 The Argument From Reciprocity

Lastly, there is the argument from *reciprocity*, which comes in both a conceptual and an empirical variety. According to the conceptual version, a being who cannot be held accountable, lacks the authority, power, or standing to hold others accountable. According to the empirical form of this argument, the intrinsic desire of moral agents to consider the claims, demands, objections, et cetera, of other agents is, as a matter of moral psychological fact, only triggered when those other agents are perceived to be mutually accountable. The empirical variety of the argument from symmetry is, in part, challenged in Paper IV. The conceptual interpretation will be addressed below.

Paper IV challenges the empirical variety of the argument from reciprocity by pointing to real-life examples of interactions where the presence and inclusion of moral patients in particular social settings seems to elicit a claimant-directed participant stance in typical adult humans. Looking at these examples, I argue, undermines the assumption that the moral motivation of justifying ourselves to others requires, or rest upon an expectation of, reciprocation.

The conceptual variety of the argument from reciprocity can be dealt with by reiterating one of the principal distinctions between contractarianism and contractualism. On the contractarian story, only those whose inclusion is mutually advantageous can be parties to the contract. This is generally taken to imply that people who cannot provide any benefits and who do not pose any threat to others' interests are excluded. In this way, reciprocity or mutuality is, for strategic or rational reasons, a central premise of the contractarian social contract. This means that many seeming moral patients, humans and nonhumans, as well as future generations, are excluded from the scope of morality set by the social contract.

According to contractualism, however, people have an intrinsic desire, essential to moral agency, to justify themselves to (and therefore consider) others. Hence, while reciprocity or mutuality are common terms in contractualist reasoning, this is not because of reasons of strategy or mutual advantage (Scanlon, 1998, p. 180). Instead, reciprocity enters the picture as a way of capturing what it means, and how it feels, to be morally obliged to someone. Rather than implying bargaining, our assumed desire to justify ourselves to others is argued to operate on the prospect of what Scanlon calls reasonable rejection (Scanlon, 1998). To respect others, simply means taking into appropriate account their evaluative perspective. This, in turn, is taken to require considering, and guiding oneself by, principles that others could not reasonably reject.

However, this does not in and by itself, justify considering moral agency-exempted beings purely as wards. Reasons to reject the reciprocity argument have been argued by Matthew Talbert to follow from central tenets of Scanlon's own theory (Talbert, 2006). Talbert believes that Scanlon's suggestion for trustees makes explicit that mutuality or reciprocity is not a requirement for moral consideration. Scanlon claims that the requirement of justifiability could be met in the case of moral agency exempted humans by considering what trustees think that they could reasonably reject to if they were able to engage in such deliberation.[82] However, if trustees can affirm the moral status of humans who allegedly lack the capacity for "judgment-sensitive" attitudes, by understanding justifiability in their case in counterfactual terms, reciprocity was never an issue to begin with.

[82] A related proposal is found in suggestions of indirect reciprocity with respect to distributive justice, for instance in the context of intergenerational justice (Page, 2006).

5.5 Concluding Remarks

This chapter has highlighted shared assumptions between practice-focused accounts of moral agency, on the one hand, and contractualist accounts of moral patiency, on the other. In Papers III and IV, I challenge the assumption that an entirely objective stance towards agents exempted from moral responsibility practices is justified. I have developed some of these arguments by relating them to the contractualist distinction between two distinct grounds of moral patiency. In short, I claim that Including or exempting others as participants in moral responsibility practices likewise involves including or exempting them as moral claimants (Paper III). Following Kantian contractualist views of moral patiency, I have argued that including or exempting someone as moral claimant can be understood as attributing one of two distinct types of moral patiency (Paper IV). Moral claimancy is typically reserved for moral patients who are seen as active participants or members of moral responsibility practices and wardship is typically attributed to moral patients who are assumed to be exempted from moral responsibility practices. Hence, understanding the difference between the claimant and ward perspectives, in turn, can serve to highlight possible harms or risks associated with taking an entirely objective stance to someone.

Using the contractualist and practice-focused literature, I extrapolate and respond to three arguments in support of the case for pure wardship. I argue that none of these arguments seem able to justify exempting moral patients as claimants. I will return to the case for a general claimant-directed participant stance and its implications for contractualism in the next and final chapter.

6 Final Discussion

The final chapter serves as a summary of the main findings from this thesis, addresses some remaining issues, and points toward potential future research questions.

6.1 Some Answers and Some Limitations

This thesis has investigated the possibility of moral agency in nonhuman entities, with artificial intelligence agents and nonhuman animals as focal cases, assuming the practice-focused approach to moral agency. The principal questions guiding the four papers and introduction have been: Can moral agency be ascribed to nonhuman entities? If so, in what sense, or to what extent, can moral agency be ascribed to them? *Should* it be so ascribed? What criteria and boundaries are valid for affirming or denying the moral agency of nonhuman beings? I will begin by recapping my suggested answers to these questions and the arguments supporting these answers.

6.1.1 Valid Criteria and Boundaries

Paper I and II make explicit a number of methodological shortcomings in both the artificial and nonhuman animal moral agency debates. The typical procedure in these debates (especially among skeptics) is to consider the possibility of nonhuman moral agency assuming a capacity-focused approach. That is, by starting from an a priori set of requirements assumed to reflect widely prevalent features in typical adult humans relevant for moral agency (3.1.1).

I have questioned the assumed prevalence of these features and their relevance as basis for requirements of moral agency. A sourcehood condition that automatically excludes an advanced programmed, albeit autonomously adapting and learning, artificial entity, runs the risk of excluding any biological organisms, such as humans, from moral agency as well. Likewise, requiring conscious reflection and deliberation for moral agency seems to set the bar of moral responsibility too high by excusing or exempting a large portion (maybe even the majority) of morally significant behavior of humans. Lastly, the emphasis on

phenomenal consciousness, especially in the artificial moral agency debate, makes explicit the importance that any purported necessary feature be specifiable in operationalizable terms.

Against this background, I followed and developed Strawson's (3.1.2) socially situated and naturalistic account of moral responsibility to suggest an alternative route. According to the practice-focused approach, the nature and requirements of moral agency should be determined by starting with the everyday contexts and behavioral patterns where this concept is assumed. Hence, assuming that moral agency designates the eligibility of an entity to have its conduct or character appraised morally, the practice-focused approach asks us to attend to practices of moral appraisal.

In addition, assuming the objective of conducting valid comparative assessments (3.3.2), it is paramount that the assumed requirements reflect typical features of the alleged paradigm target: the moral responsibility practices of typical adult humans. As such, requirements need to be applicable to, and accurately represent, everyday moral interactions and the features required for engaging in them.

I suggested a modest empirically informed practice-focused account (3.4). Moral agency can be understood as a social-normative competence. Moral agents acquire, and are able to conform to, socially mediated standards of appropriate (right/good) conduct and to refrain from conduct that is inappropriate (wrong/bad) on these standards. Importantly, a moral agent is sensitive and responsive to the morally relevant features of a situation. Such features may pertain to harm, wellbeing, fairness, rights, interests, et cetera, of oneself or others.

Westra & Andrews write: "Whether or not a pattern of behavioral conformity counts as a normative regularity depends on how members of a given community respond to individual cases of conformity and nonconformity." (2022, p. 10). Hence, one can infer behavioral prescriptions and prohibitions by looking for whether a behavior elicits (positive or negative) social responses. A central aspect of moral agency is liability to reactive attitudes. This requires being prone to, as well as recognizing and responding to, certain expressive emotions, such as *moral anger* (Paper III), guilt, or gratitude, which convey appraisals of rightness/wrongness, goodness/badness, appropriateness/inappropriateness, and respectfulness/offensiveness.

That is, moral agency requires the competence to internalize, recognize, and be moved by normative considerations pertaining to how one behaves and sees others

as well as the competence to react to cases of (non)conformity by engaging in practices of social maintenance.

However, a good account of moral agency needs to also provide some guidance regarding when it is appropriate to include or exempt an agent from moral responsibility practices. As previously mentioned (3.3.2 and 4.1.7), this creates a tension between descriptive desiderata on the one hand (that is, how and why we in fact include or exempt) and normative desiderata on the other (that is, how and why we ought to include or exempt) (see Argetsinger & Vargas, 2022).

6.1.2 Normative Guidance

How can the goal of deriving prescriptions be reconciled with the objective of accurately representing how people actually function and behave? I believe that part of a solution is to be found in fittingness accounts of moral responsibility ascription. A promising such account is discussed and defended in Paper III. The communicative emotion account of blame identifies blame with an expressive emotion episode (Macnamara, 2015a). According to this view, holding other agents responsible essentially amounts to having a particular negatively valenced object-directed expressive emotion.

Reactive attitudes, such as resentment and indignation, are types of moral anger that appraise an agent's conduct or attitude as wrong, bad, offensive or seriously inappropriate. What is more, blaming reactions involve characteristic action tendencies functionally aimed at conveying blame's content to the perceived transgressor. Blame *seeks* a response in the target or recipient.

As such, the fittingness of blame can be taken to depend on two connected conditions. Firstly, moral anger is fitting to the extent that its appraisal of someone's conduct or attitude is correct, that is, that the evaluation of the action or attitude as a blameworthy violation or transgression fits the action or attitude. In other words, blame is fitting if its object *merits* blame (Shoemaker, 2017).

Secondly, the expression of blame is appropriate depending on the assumed function of directed blame. Following the response-seeking communicative understanding, blame is appropriate to the extent that the assumed transgressor or other recipient (such as onlookers) are capable of uptake of blame's evaluative content.

Such uptake, however, can be understood in a variety of ways. I will compare what I take to be two popular, mutually non-exclusive, definitions. First, an agent may be capable of uptake if they already internalized the moral consideration in

question but had to be reminded about it. That is, circumstances or other considerations may have blocked or dulled the agent's sensitivity and responsiveness to the consideration. Second, another way of conceiving of uptake is to say that the agent can take on new moral considerations. Hence, according to this latter definition, blame may be appropriate if the recipient has the general capacity to be sensitized to moral considerations.

Of course, the distinctiveness of these conceptions depends, in part, on how one specifies *reminding* as opposed to sensitizing to a new consideration. Some would say that as long as the moral consideration in question is something that follows from a coherent set of values it is not, in fact new.[83] An agent can act wrongly "by her own lights" even if the knowledge in question "was not at the forefront of the mind at that moment" (Mason, 2019, p. 103).

An alternative reading, in line with the wider definition of uptake, is to say that an agent's eligibility for moral address is contingent on whether they are capable of being sensitized to the consideration in question. Hence, the agent does not need to have acted wrong by their own lights for reactions or ascriptions of moral responsibility to be appropriate. It is sufficient that they can come to appreciate and adjust to the moral consideration. This wider notion of uptake reflects the view of contemporary instrumentalist accounts of moral responsibility and agency. I will return to such accounts further down (6.2).

6.1.3 Attributing Moral Agency to Nonhuman Entities

I have defined the nature and requirements of moral agency in terms of a social-normative competence to internalize social norms or standards pertaining to moral considerations broadly construed, and to engage in social maintenance behaviors assuming such norms or standards. Any entity who displays these competencies is a moral agent. Hence, I have argued that moral agency can be attributed to some existing nonhuman animals. I also believe that artificial moral agents are, in principle, possible.

Canids, like dogs seem to fully meet the suggested requirements of moral agency in relation to conspecifics. Dogs acquire social norms by recognizing and internalizing standards of appropriate behavior, some of which pertain to welfare, harm, fairness, respect, et cetera. The rules of social play, for instance, are

[83] For example, Mason (2019) argues that "ordinary blame" is to judge that someone "has failed by her own lights" (2019, p. 103). As such, the warrant of ordinary blame presupposes that the perceived wrongdoing is an act of "subjective wrongdoing" (2019, p. 103).

standards of appropriate conduct where rightness and wrongness of a situation pertains to concerns such as harm, fairness, and honesty. Dogs also engage in practices of social maintenance by reacting to perceived violations of these rules and by responding to such calls or invitations. Importantly, moral norms and moral agency are made visible through such exchanges.

However, nonhuman animals may not be equally eligible for moral address by humans. In Paper III, I argued that differences in language and asymmetries in scaffolding resources make it much more difficult or costly for some agents than others to be sensitized and respond appropriately to (some) moral considerations.

Hence, asymmetries in various types of communicative skills, as well as social and material resources, can render some agents, practically speaking, exempt from directed blame. Differences in social context and (moral) psychological make-up can serve to exempt, say, dogs as well as young children from many (if not most) norms or expectations pertaining to typical adult humans.

However, while toddlers and dogs, for example, may be "off the hook" (Delon, in press, p. 17) of most standards assumed to apply to typical adult humans, they can still hold others morally responsible for perceived violations or transgressions. For example, I have argued that dogs seem to qualify as cross-species participants by being able to morally address perceived moral violations in humans. As such, humans may have reason to be attentive and responsive to potential instances of nonhuman moral address.

While machines do not seem to currently exhibit the mentioned required competencies as a whole, I believe that there could be artificial moral agents in the near future. For instance, some chatbots already appear to be able to respond quite competently to calls or demands to provide reasons for their conduct. And human users are sometimes called on to justify or take responsibility for their behavior by chatbots. As such, some AI could be thought of as qualifying as participants in moral responsibility practices in the answerability-sense. While such eligibility is but one aspect of moral agency, the explicitness of linguistic interaction may be sufficient to elicit the ascription of more fully-fledged moral agency by human users.

Importantly, this thesis has argued that the development and deployment of increasingly advanced, autonomous, and socially competent AI applications, calls for a *normative*, rather than descriptive, approach to the question of artificial moral agency (Paper III). The increasing reliance on AI-tools for various types of decision-making processes, highlights the shortcomings of traditional intra-individual, static and monistic moral agency accounts in favor of accounts that are

able to accommodate the contextual, interactive, and dynamic nature of moral agency attribution.

Policymakers, developers and users need to consider the implications of machines being included in social practices where moral agency is assumed. Such consideration may, for instance, include questions about how the appearance and behavior of an AI application may affect moral agency attribution. However, including machines in practices of moral responsibility may also be assessed with a view to the further effects of such attribution. For example, if some artificial entity appears in such a way that its behavior elicits moral appraisal, how does this impact the possibility to assess and determine the moral responsibility of human developers and users? What is more, given a moral psychological link between attributions of moral agency and moral patiency (Chapter 5 and Paper IV) the inclusion of AI in moral responsibility practices may lead to the ascription of moral patiency to artificial entities.

6.1.4 Limitations

Needless to say, there are some limitations to the practice-focused approach assumed and defended here. Since this approach is concerned with inter-relational communicative practices, it may be argued to leave out important aspects of moral responsibility and agency, for example, ascriptions of moral responsibility that are covert or private. Hence, assuming that holding others responsible, through reactive attitudes, can be unexpressed (that is, private), an expressive or communicative understanding of moral agency and responsibility would seem unable to account for such states. A possible implication could be that the conception of moral agency defended in this thesis is incomplete.

It is true that the practice-focused approach defended in this thesis assumes that the reactive attitudes are inherently communicative and paradigmatically expressed. However, that does not mean that the account cannot recognize and account for private blame, for example. Nor does it undermine the attribution of moral claimancy in nonstandard cases, for instance to dogs or toddlers, on the assumption that such agents cannot entertain private blame. This is because the internalist stance of the communicative emotion account does not preclude the possibility of unexpressed blame. While, moral anger, for instance, necessarily involves an action priming element, its disposing effect does not exist in isolation from other behavioral pulls or considerations.

For example, a human or a dog may feel resentment or indignation toward someone, and consequently be subject to the action priming effects of moral anger, but still not express or behave in ways conveying that anger. They may, for instance, be afraid of the consequences were they to voice their blame towards a particular agent or in a particular context. Or they might inhibit their inclination to confront the perceived violator due to external considerations, like fatigue. However, just like an unsent letter, private blame still *contains a message*, and still is *supposed to* evoke uptake of its content (Macnamara, 2015a).[84] I therefore maintain that a practice-focused and communicative understanding of participation in moral responsibility practices can accommodate private blame.

Another possible objection toward the communicative emotion account follows from a cognitivist conception of blame, which conceives of (some) blame as mere judgments of responsibility. Assuming such an account, the practice-focused view does not capture all types of responsibility ascription, but only a subclass of them. I concede that one can distinguish a sense of moral responsibility ascription that is void of the characteristic affective or conative elements of, say, moral anger (Paper III) or reactive attitudes. However, I believe that judging responsible in this cool and detached way is "conceptually distinct" (Mason, 2019, p. 100) from *holding* responsible.

One may also note that the fittingness conception is not an account of desert. For instance, it cannot provide an answer to why and when an agent *deserves* to be blamed in the traditional backward-looking sense. Hence, conceiving of the appropriateness of moral responsibility ascription in terms of fittingness may, on the desert account, be considered to miss the mark. While a proper treatment of this question is beyond the scope of this thesis, I believe that considerations about desert can be compatible with the communicative emotion account. For example, while the appropriateness of blame depends on questions about fittingness and the possibility of uptake, the warrant of the shape or form of the *particular blame response* depends on various further factors (see Arpaly, 2002).

In addition, it is worth noting that I take ascribing moral responsibility and punishing to be different things. This does not rule out the possibility that blame may many times feel like punishment for the recipient, or that blaming responses may often co-occur with, or be followed by, punishment. But on the communicative emotion account, blame and punishment are distinct.

---

[84] Of course, as stated elsewhere (Paper III, and Chapter 4, 4.1.1 and 4.1.2), practices of moral appraisal may have other functions and values.

6.2 Instrumentalist Considerations

I believe that the conclusions and arguments developed in this thesis can serve to contribute to a wider trend in the literature on moral responsibility and agency. Recent years have seen a renaissance of sorts for accounts that understand the justification of moral responsibility practices "in instrumentalist terms" (Vargas, 2022, p. 3; see also Vargas, 2013; McGeer & Pettit, 2015; Jefferson, 2019). A core assumption of these *moral influence, agency cultivating* or *moral scaffolding* accounts, and an underlying premise of Paper IV, is that "[w]e can "train up" cognition for particular environments. Environments can foster particular patterns of cares and commitments that shape what agents perceive as reasons. We can also restructure our environments to better exploit our cognitive and affective dispositions, and to better express and realize our cares and commitments." (Vargas, 2018, pp. 10-11).

Considerations about nonstandard moral agency can serve to make explicit, clarify, and improve, instrumentalist views. I have argued that the objective of making *valid* comparative assessment of moral agency, supports the case for a modest empirically informed account of moral agency. Rather than being a primarily intra-individual robust capacity, sensitivity and responsiveness to moral considerations is instead argued to be an environmentally and socially contingent competence.

If the "agential capacities" of typical adult humans are "inescapably vulnerable, for better or worse, to the dynamics of social interaction" (Mackenzie, 2018, pp. 76-77), this speaks against certain and neat demarcations between those who are included and those who are exempt from moral responsibility practices. Likewise, if even paradigm moral agents "rely on others to attune and calibrate our tracking of moral considerations" (Vargas, 2018, p. 128), the inclusion or exemption of nonstandard cases in moral responsibility practices may have normative implications.

In Paper IV I argue that a key, albeit overlooked, source of moral scaffolding can be found by considering the evaluative perspective of moral patients that are typically exempted from moral agency. Viewing all moral patients as moral claimant participants is therefore argued to scaffold the sensitivity and responsiveness of the stance-taker by widening the scope of their perceived moral "audience" (McGeer & Pettit, 2015, p. 169).

Moreover, the assumed primacy of our *moral audience* as a key source of moral scaffolding, seems to reflect a central contractualist premise, namely, that moral motivation (partly) consists in the noninstrumental desire to be able to justify one's

actions to others (Scanlon, 1982/2013). As such, the normative case made for a general claimant-directed participant stance in Paper IV may provide possible solutions to a long-standing problem with social contract accounts that link moral patiency to moral agency.

As previously discussed, Kantian contractualist theories cannot fully accommodate all beings who we seem to have independent, noninstrumental reasons to recognize as moral patients (Chapter 5). The point is not that contractualist theories entirely fail to account for moral agency exempted sentient beings. As mentioned, there are various suggestions as to how one may morally consider some moral patients despite recognizing their alleged absence from the moral deliberative space, for instance, by considering them purely in terms of welfare considerations, or by appointing trustees with the task of considering their best interest.

However, none of these suggestions seem able to address the fundamental issue at stake, namely, that Kantian and contractualist accounts, typically conceived, assume that moral agency-exempted moral patients can and should at most be considered only as passive wards. However, this passive, indirect, derivative way of accounting for sentient beings offers a mode of moral inquiry that, on its own, is inadequate and potentially reckless and harmful (Paper IV).

Following a loosely Scanlonian (1982/2013) formulation of this thesis, a revision of the moral psychological thesis might look something like the following: Moral motivation (partly) consists in the noninstrumental desire to be able to justify one's actions to all beings who could pose actual, prospective, or hypothetical objections. Actions are not perceived to be wrong because they cannot be justified to mutually accountable moral agents. Instead, actions are perceived to be wrong because they cannot be justified to all morally significant beings (with a vested interest in having a say about said action).

Needless to say, it may be the case that humans are disposed in ways that make the adoption of the claimant perspective to small children and nonhuman animals very difficult or even impossible. However, the feasibility of this stance to typically moral agency exempt beings may also depend on the particular shape or strategy of intervention. While I am less optimistic about the prospect of interventions at the level of individual agents, I believe that the real-world cases described in Paper IV (sec. 4.2) speak in favor of environmental interventions. Ultimately, however, the possibility of a general claimant stance to moral patients is an empirical question. This brings us to the last section of this final chapter.

6.3 Questions for Future Research

Assuming that the claimant/ward-distinction captures a tendency in humans to consider a moral patient as either one of these positions at a time, it may contribute to discussions and research in moral psychology, in particular research on mind perception and moral status attribution (see Hallgren, 2012; Robbins & Jack, 2006). Assuming that people attribute wardship or claimancy as different types of moral status, this would call into question the dichotomous view of moral agent/moral patient in *moral typecasting* theory (Gray et al., 2012) by showing that moral agency attribution, in the form of seeing someone as a moral claimant, involves a distinct other-regarding perspective. The claimancy/ward distinction may also add to the experience/agency distinction of the *two-source hypothesis* (Sytsma & Machery, 2012) by suggesting a finer distinction within the agency source of moral standing. Seeing someone as a moral claimant would then constitute an additional way of viewing others as minded and morally significant beings, above and beyond seeing them as cognitively complex.

An additional promising future inquiry would be to analyze accounts and conclusions of moral agency more seriously and explicitly in light of *levels of analysis* (MacDougall-Shackleton, 2011).[85] I believe that doing so may serve to highlight the extent to which putatively conflicting philosophical theories and arguments of moral responsibility and agency are non-exclusionary. I am here not merely implying that different theoretical objectives, such as suggestions of the nature, structure *or* justification of moral responsibility practices (see Vargas, 2022) may render apparent distinct accounts compatible. Rather, I suggest that careful attention to putatively conflicting accounts or explanations of the same target may prove to constitute a "false debate" (MacDougall-Shackleton, 2011, p. 2077).[86] The point here is that explanations at different levels of analysis or inquiry are, per definition, non-exclusive.

The ability and inclination of an agent (whether human or nonhuman animal) to, for example, express resentment can simultaneously be understood in terms of

---

[85] For instance, Tinbergen's four questions (1963; see also Sherman, 1988), and Mayr's (1961) distinction between ultimate and proximate levels, suggest that we can explain behavior by way of various complementary levels of explanation. Following the latter, behavior can be explained in terms of both mechanistic causes (physiological, cognitive, etc.) and adaptive functions. An additional way of talking about levels of analysis is found in levels of reductionism, such as molecular, genetic, physiological and behavioral levels of inquiry (see also MacDougall-Shackleton, 2011).

[86] See also Macnamara (2015b, sec. 4) for a discussion about whether seemingly conflicting views on blame really are incompatible.

its adaptive function(s), such as the ecological and intraspecies benefits of the behavior, *and* in terms of the physiological and psychological mechanisms responsible for it, such as its underlying neural, cognitive, and emotional processes or states. Hence, careful analysis of reductionist types of objections against claims of putative *moral behavior* in humans and nonhumans may prove to not necessarily undermine such suggestions.

Lastly, the communicative nature of (some) moral responsibility practices not only underscores the potential but also emphasizes the advantages of conducting studies that explore social interactions between humans and machines, as well as humans and animals. Studies could shed light on the emergence or prevalence of, for example, shared norms, standards, or rules. A promising type of such interaction may, for example, be found in joint or collaborative endeavors, like games. The proposed moral-psychological connection between ascriptions of moral agency and patiency could be further explored and evaluated through these studies. For instance, researchers could examine an agent's moral intuitions and decisions concerning a moral patient, assessing whether these judgments and choices vary based on the extent and nature of the agent's interaction with the patient.

7 Paper Summaries

Paper I: A Normative Approach to Artificial Moral Agency

In this paper, Christian Munthe and I ask: What conditions are sufficient and necessary for an entity to be a moral agent? We address this question by reviewing different positions in the artificial moral agency (or AMA) debate. Despite the fact that these discussions revolve around questions of moral agency in *machines*, we can identify two main rivalling conceptions of *human* moral agency. We dub these positions the *standard view* and the *functionalist view*, respectively. Although there are variations within each of these views, they typically diverge on two issues. First, they disagree on the importance of phenomenal consciousness for moral agency. Second, they disagree on the possibility of AMA given diverging assumptions about independence, autonomy, or freedom.

According to the standard view, an entity needs to have conscious mental states in order to be able to act, and not just behave. This is because phenomenal consciousness is necessary for an entity to engage in the rational decision-making and evaluative processes required for moral agency. An additional argument put forth for the importance of consciousness is that subjective mental states, such as guilt or remorse, are necessary for ascriptions of moral responsibility to be meaningful in the first place. It would, for instance, not make sense to blame a machine, and hold it responsible, if it were not capable of having moral emotions such as those mentioned.

Functionalists, on the other hand, deny the relevance of subjective mental states for moral agency. They point to the fact that we depend on observable features when ascribing conscious states even to typical adult humans, the supposed paradigm of moral agents. Given that a good theory moral agency should preserve current practices of ascribing moral agency, as well as the assumption that humans are moral agents, functionalists claim that a separate condition for consciousness would fail to accommodate that. Instead, we should require only those observable features normally taken to be required for moral agency. This, they claim, is in line with the assumptions and practices already in use when we

ascribe consciousness and moral agency to humans, and therefore lends support to the functionalist, as opposed to the standard, view of moral agency.

A related suggestion to the one above, is found in what we choose to call the *epistemic argument*. Starting from our epistemic practices of identifying mental states and ascribing moral agency, proponents of this argument propose a pragmatic solution to the issue of consciousness for AMA. The suggestion is that we should determine moral agency solely based on whether an entity meets certain observable criteria assumed to indicate the possession of consciousness. In this way, the epistemic argument avoids conceptual reform by maintaining the standard requirement for subjective mental states, while still accommodating for the epistemic objection put forth by functionalists and others.

The second issue separating the standard and functional view, is their position on the possibility of AMA meeting a condition of independence. According to the standard interpretation of this condition, AMA is, in principle, impossible because artificial entities lack the right kind of autonomy or freedom required for moral agency. Robots and computers, however advanced and autonomous, are merely behaving on basis of their programming. Just like any other tool or prosthetic, they are merely extensions of human intentionality. As such, they cannot be ascribed ownership for what they do or how they function and fail to meet the source control condition for moral agency.

Functionalists, refute this conclusion, and argue that machines need not be construed in a way that makes them predictable or wholly reliant on original programming. Just like humans, artificial entities may be designed to adapt and evolve in response to external cues. For instance, machines may be equipped such that they are able to modify their programming. An artificial entity designed to behave in accordance with a set of normative rules, could then adjust those rules similar to how humans learn and adapt to changing normative environments. Furthermore, like machines, humans can be said to be products of code and programming in virtue of inheriting a genetic blueprint from their parents and being informed and instructed by upbringing and later environmental input. Excluding machines because they don't meet the source control condition, seems to fail to discriminate between humans and artificial entities.

Our evaluation of the AMA debate gives rise to three main conclusions. Our first conclusion is that the frequently assumed importance of consciousness for moral agency, is questionable. Subjective mental states are mostly motivated on their relevance for other features, such as rationality, moral competence, the right kind of independence and moral responsibility, all of which can be understood in

dispositional terms. Certain notions of moral competence may however still motivate consciousness for moral agency. As may a requirement for subjective mental states based on an assumed link between moral agency and moral patiency. However, we concur with the epistemic argument that whatever features taken to indicate consciousness (or any other required feature) in humans, should suffice for the ascription of it in machines.

Our second conclusion is that there appears to be confusion about key concepts, and their relations, in the AMA debate. For instance, the term autonomy is used in highly diverging ways, signifying everything from the ability to move without direct human control, to an advanced capacity for independent decision-making. There is also disagreement as well as diverging understanding of the importance and meaning of moral rationality and moral competence. If moral agency requires things like a moral sense, moral intuition, or phronesis, it is not clear why such competence or sensibility would exclude artificial entities from moral agency, as such features or capacities can be understood in dispositional terms. Furthermore, despite the philosophically widely held assumption that moral agency makes an entity eligible for ascriptions of moral responsibility, the relationship between these two concepts is not as clear-cut in the AMA debate. These factors make it unclear if, and to what extent, seemingly different positions really are in conflict.

Our third conclusion is that the central disagreement between standard and functionalist conceptions of moral agency offers little help with how to approach artificial entities in practice. This, despite an ever-increasing need for straightforward guidelines on how to relate to, and treat, artificial entities in contexts and practices that involve moral agency. To meet this need, we suggest a methodological re-direction of the AMA debate. We propose shifting from the predominant theoretical focus to a straightforwardly normative approach. We ask to what extent, and how, artificial entities should be included in human practices where moral agency is typically assumed. This normative approach actualizes questions about the sharing of agency and responsibility with machines, issues about safety and effectiveness, concerns about possible effects on human (moral) psychology, and questions about the possibility of artificial moral patients.

Paper II: A Practice-Focused Case for Animal Moral Agency

In this paper, I suggest an answer to the question asked in the previous paper, by proposing an alternative approach to what I call the *capacity-focused* approach to moral agency. According to this widely embraced approach, moral agency requires certain intra-personal capacities, processes, or states, such as consciousness or rational deliberation. This type of approach is represented in both the standard, as well as the functional, views of moral agency, outlined in Paper I. In this paper I suggest shifting from these types of approaches to one where moral agency is viewed as something primarily inter-relational and dispositional.

According to this *practice-focused* notion, moral agency is understood as the participation in certain social practices where ascriptions of moral responsibility are held, expressed, and undertaken. I aim to show that these moral responsibility practices (or MRPs) are prevalent in other animals, such as canids, like dogs and wolves. The practice-focused approach to moral agency thus seems to make the prospect of non-human moral agency more probable than commonly thought. To demonstrate the soundness and validity of this argument, I answer the following three questions: What are the main features of MRPs? What does it take to participate in them? Are there any credible analogs to MRP participation in nonhuman animals? The two first questions are answered by analyzing various practice-focused accounts of (human) moral agency. The third question is answered by presenting empirical data of canid social play and cognition, along with an additional, bolstering, argument.

According to practice-focused approaches to (human) moral agency, MRPs and MRP participation can be summarized as follows: we share strong dispositions to recognize, internalize and enforce social norms. These dispositions are reflected in the practices surrounding indicative of social norms, such as certain moral attitudes and expressive behaviors meant to communicate or signal how well someone's action or omission aligns with our expectations.

Blame, praise, and other forms of moral reactions can be expressed verbally, as well as nonverbally. The target of such address can respond with attitudes or behaviors, such as remorse, asking for forgiveness, or by providing justifications or explanations. Being a participant of MRPs means that one engages others, and is engaged with, in moral exchanges. One may sometimes be excused from blaming and praising reactions if circumstances explain away the wrongness of the perceived transgression. Some people, like young children and adults with agency-

undermining conditions, may be exempted from ordinary MRP participation in virtue of not being eligible for moral reactions in the first place. This is because participation requires practice-relevant dispositions, such as recognizing and internalizing norms, being able to react to transgressions and to understand, and respond to, moral address.

Canid social play is a good example of a nonhuman animal interaction where there are clear expectations of how one should behave and where transgressions are met with certain, well-defined, reaction-response exchanges. Social play is initiated upon a play invitation, communicated through, for instance, a play bow. Here, the dog or wolf, will crouch down on her front legs, facing the prospective play partner, with her back upright, and tail wagging. If the other canid accepts the invitation, play will commence and may consist in interactions such as wrestling, chasing, etc. The play bow, and other play signals, appear to function as punctuation or modifiers throughout the play interaction, communicating that seemingly aggressive or easily misinterpreted behavior is, in fact, benevolent.

Canids appear to adhere to certain rules or normative expectations regarding social play. When the rules of play are transgressed, canids will react by pausing, cocking their head, and squinting, as if asking why the other party, for example, bit too hard, refused to switch roles or violated the standards of play in other ways. Sometimes, the reaction to a perceived transgression might be more direct, like when one canid growls or air snaps at the perpetrator. Stronger reactions, like these, have however typically been preceded by several subtle or polite requests and reminders. The alleged transgressor may respond to such reactions by performing a play bow, ensuring her playmate that she is still just playing. And usually, transgressions are forgiven and forgotten. Repeated violators, who often break the rules of play, will however be chased off and avoided. Puppies are generally treated mildly when breaking the rules of social play.

Canid social play thus appears to involve the features of human MRP participation described earlier. When canids abide by the rules of social plays the result is behaviors that imply trust or praise, such as affiliation, continued play, role-reversal, and self-handicapping. Surprise behaviors or warnings, such as growls, from one play-partner are forms of moral address, signaling that the behavior did not meet her expectations. Play signaling or reconciliatory behavior, may save the situation by communicating that the perceived transgressor is sorry, just wanted to play, or acknowledges the reaction.

Puppies or juveniles are wholly, or partly exempted, from when appearing to offend the rules of play, similarly to how young children, and others, are not

considered to be (full) participants of MRPs. Hence, the elements of MRPs seem to be prevalent in canids, and some canids appear to meet the requirements of participation. Some canids thus are moral agents in virtue of being participants of MRPs.

However, to show that canid social play behavior is similar in a relevant, rather than merely apparent way, to human MRPs, an additional argument is needed. I support the analogy with what I call The Function Argument. According to this argument, human MRPs and canid social play behavior are both examples of behaviors with the function to promote and sustain peace and cooperation. Furthermore, the function of these behaviors and practices is realized through shared proximate mechanisms: being able to recognize social norms, internalize them, communicate about perceived transgressions, and respond and adjust accordingly.

A further objection against the analogy argument is to claim that the described reactions and responses of canids, do not constitute relevant analogs to human moral practices of asking for, and giving, reasons. Moral address, like blame or criticism, and moral responses, like excuses, explanations, or justifications, are central elements of human MRPs. If these elements are missing in canid social practices, the latter cannot be considered relevant analogs to human MRPs.

This challenge, I argue, does not undermine the analogy. Although the paradigm example of a human moral exchange is typically portrayed in terms of linguistic modes of communication, many day-to-day moral reactions and responses rely on non-linguistic modes, such as facial expression, gaze, posture, and non-linguistic vocalizations. An angry stare from someone usually prompts us to respond by figuring out whether we may have harmed or offended them. If so, we are disposed to acknowledge our transgression and communicate that we didn't mean to, or that we are sorry. Such responses can be conveyed through mere facial expressions and gestures. In a similar sense, dogs and wolves utilize movement, posture, vocalizations, and facial expressions to address transgressions, and respond to such address. Canid norm exchanges are not fixed or mechanic, but sensitive to, not only the norms or rules, but to contextual factors such as the other party's attention, emotional state, age, perceptual access, and, not least, their response.

However, one may still object to the arguments provided here, by criticizing the practice-focused approach itself. Even if canids were to behave in accordance with moral norms, this may not be sufficient for ascribing them moral agency. For a view of moral agency and responsibility to be justified, it needs to also provide

an account of why and when an entity's actions are assessable as good/bad or right/wrong. Many capacity-focused views do this by arguing that the character, beliefs, actions, or omissions of moral agents need to arise in the right way (in terms of the right kind of processes or contents). Entities who, for instance, lack the right kind, or sufficient degree, of control or knowledge, cannot be ascribed moral responsibility, and thus are not blame- or praiseworthy.

Even so, the practice-focused approach is compatible with normative accounts. Many practice-focused theories of human moral agency are, for instance, paired with forward-looking or moral influence-ideas, and others with views about attributability or virtue. Although this does not close the case, as these accounts may themselves be questioned, the plausibility of any requirement for moral agency will depend on the validity of the underlying claims. Various capacity-focused requirements, such as conscious deliberation and awareness of motivations and intentions, are challenged by recent empirical findings in psychology and cognition. I believe that this constitutes an independent reason for re-evaluating our pre-theoretical intuitions on moral agency, as well as for asking how we should think about, use, and assess standards for moral agency. Furthermore, I think that the dispositional, inter-relational perspective inherent to the practice-focused approach fares well in light of these findings.

Paper III: The Moral Claimant Account of Moral Agency

In this paper, I highlight the implications of an often-overlooked aspect of participation in moral responsibility practices for the possibility of moral agency in non-paradigm entities. Both capacity-focused and practice-focused approaches to moral agency typically characterize moral agency in terms of eligibility for being the object or target of moral assessments and reactions. Likewise, various debates about the boundaries of moral agency have often assumed that the key question is whether a certain entity qualifies for ascriptions of moral responsibility.

I question this common assumption and argue that moral agency, understood as participation in moral responsibility practices, encompasses more than merely participating as the recipient or target of moral appraisal, that is, as a moral defendant. One may also participate as the source or maker of moral address, that is, as a moral claimant. More importantly, I argue that an agent can be a moral claimant despite not being eligible as a moral defendant. In this way, participating

as a moral claimant constitutes a distinct form of participation in moral responsibility practices, and expands the theoretical room of moral agency.

To clarify the centrality of the claimant aspect of moral agency, I turn to communicative varieties of the practice-focused approach. According to these accounts, the requirements for moral agency are to be understood by appeal to *moral address* (Watson, 1987/2004). When we take the participant stance to someone and react to their conduct, we do so on the assumption that they will understand the message we are trying to convey. In this sense, the participant stance assumes that the other party can be engaged in a moral exchange.

However, when looking at moral agency as the participation in certain communicative practices, it becomes obvious that there are, in fact, two positions, or roles, involved. A moral conversation implies (at least) two parties. One is the addressing party, and the other is the addressed party. Shifting perspectives like this illuminates that the notion of moral agency as a status or set of features that makes one eligible for moral reactions overlooks a central aspect of moral agency. Participation in moral responsibility practices also involves the addressing of others, and the eligibility requirements of this role may differ from those of being eligible as a defendant. This may have implications for the scope of possible moral agents and for everyday moral practices.

Considering encounters with typically exempted beings, such as young children and dogs, are examples of interactions where they seem to occupy the moral claimant role. For instance, young children may make moral demands or claims on us, despite being ineligible for most (if not all) moral reactions themselves. A toddler can react in ways indistinguishable from moral resentment when we fail to deliver on a promise. Just imagine the look and expression of a four-year old who angrily states that you broke your promise to them.

Similarly, nonhuman animals, like dogs, seem to engage with us in ways, and about matters, that appear to fit the moral claimant account of MRP participation. For example, a dog may make use of polite and gentle means of stating "here but not closer" toward a pushy visitor. If these reminders and demands are ignored, however, she may very well turn to more forceful means of communication, such as growls, barks, or air snaps. Despite uncertainty about the eligibility of these beings as defendants, they do seem to fit the conception of moral address as involving certain demands or claims, communicated as such. In this way, the participant stance should be understood as consisting of two distinct stances: the *defendant-directed* participant stance and the *claimant-directed* participant stance. Each of which track and imply distinct participatory positions.

However, seemingly angry reactions toward someone may not suffice as instances of moral address. One can, for instance, get angry at inanimate objects, like a malfunctioning piece of electronics. Angry and directed expressions are thus not, in and by themselves, sufficient for an interaction to count as a moral exchange. I therefore suggest requirements for moral claimant eligibility given a particular theory of moral address, like blame. According to the communicative emotion account, moral reactions are distinct from other kinds of reactions by virtue of their communicative content and function. Moral anger plays a specific role in eliciting a certain response in the recipient. Being a moral claimant means that one is disposed to react to perceived transgressions with certain angry emotion episodes involving certain sensations and feelings, appraisals of slights or violations, and expressive behaviors and action tendencies conveying this content to the recipient. Feeling resentment toward someone, thus means that we have an emotional state of having been wronged by them. And these attitudes are accompanied by certain facial features and other expressive behaviors that communicate this representational content to the target.

To be an eligible target of such address requires that one is capable of responding in the right way to moral address by, for instance, recognizing the content, expressing remorse or guilt, or by providing explanations or justifications. A more long-term type of response would be to adjust one's behavior in according to the demands perceived in the moral address. The general suitability of the defendant-directed participant stance, therefore, depends on whether uptake of the right kind is possible. Based on this, many typically exempted beings, do seem to qualify as moral claimants. Young children and dogs, for example, seem capable of having feelings of heat or anger, appraisals of having been offended or wronged, and presenting the offensive action or agent as something to confront.

If being a moral defendant means that one is eligible for moral address, what kind of eligibility is implied in moral claimant eligibility? To answer this question, one needs to attend to the function of moral response. Because as the general suitability of the defendant-directed participant stance depends on the function of moral address, so does the suitability of the claimant-directed participant stance depend on the functional aim of moral response. Moral response can consist of an explanation, excuse, or even a justification. However, the general aim of moral response is to express moral uptake, that is, to signal or communicate that one has, indeed, *received the message*. In this sense, moral response is the salient or expressive aspect of moral uptake. The general suitability of moral response depends on whether the recipient can recognize such response.

The claimant-directed participant stance is a perspective from which one is disposed to morally respond to the moral address of others. A being is a moral claimant by being a suitable recipient of moral response. There are forms of moral response that young children and dogs anticipate and recognize. For example, even toddlers can understand and adjust their attitude to expressions of remorse or spoken utterances conveying that one is sorry or that one did not mean to. Likewise, dogs use and recognize reconciliatory expressions with conspecifics, and studies show that they use appeasement behaviors in interactions with humans. It seems possible that one can intelligibly respond to the demands of a dog by, for instance, providing them space and talking to them in a soft voice.

I argue that moral address and moral response have distinct functional roles. The function of moral address is to evoke qualified uptake of its content, while the function of moral response is to, among other things, express, signal, or communicate uptake. Although the claimant and defendant participatory roles imply the same basic features, their corresponding eligibilities may show up asymmetrically *between* participants. For example, given the uneven distribution of psychological, social and material resources among participants, there are some beings who may not be eligible as recipients of moral address in relation to some other participants, but who may still qualify as sources or senders of such address. As such, these beings are *claimant-heavy* participants in moral responsibility practices.

Paper IV: Moral Patiency Grounds Partial Moral Agency

In this paper, I argue that, although moral patiency and moral agency are distinct concepts, we have normative reasons to regard the former as a ground for a partial form of the latter. This means that beings who are moral patients but who do not meet requirements for participation in moral responsibility practices, still qualify for a certain type of recognition and treatment typically involved in our recognition and treatment of moral agents.

It is widely assumed that moral patients can be accounted for qua moral patients despite being denied moral agency. Assuming a practice-focused approach to moral agency, the participant stance is thought to be warranted to participants of moral responsibility practices. Beings who are wholly exempt from participation are accounted for by the objective stance. From this perspective, they are denied moral agency but may still be recognized as moral patients. Young children,

nonhuman animals and adults with alleged moral agency-undermining conditions or features are thus considered to morally matter in their own right, despite being denied moral agency.

In this paper, I question this assumption and argue that a wholly objective stance is, practically speaking, often detrimental to one's sensitivity and responsiveness to moral considerations regarding the exempted party. This is so, I argue, because the participant stance involves a valuable other-regarding perspective, unavailable from a wholly objective stance. The recognition of someone as a claimant participant disposes the stance-taker in moral psychologically favorable ways. Seeing the other party as source or maker of moral claims and demands, as opposed to merely an object of benevolent concern, disposes one to perceive a wider range of morally relevant facts and considerations about them. Conversely, merely seeing someone as a *ward*, can limit or distort one's moral sensitivity. Hence, given a commitment to promote sensitivity and responsiveness to moral considerations, we seem to have normative reasons to add the claimant-directed participant stance to the way we see and regard moral patients.

An increasingly popular way of explaining and justifying moral responsibility practices is to appeal to their suitability for *cultivating* agency (Vargas, 2013). Blame, moral criticism, and moral feedback are justified because they are responses that can scaffold and promote sensitivity and responsiveness to moral considerations. To ensure that moral responsibility reactions scaffold, instead of impede, moral agency, one needs to assess and evaluate the effect of various environments, practices, and dispositions. In this paper, I am interested in assessing the suitability of our exempting practices.

Assuming a practice-focused approach to moral agency, moral agents are participants in moral responsibility practices. As such, the participant stance is a perspective we take to beings who engage, and are engageable with, in such practices. However, large populations of beings, such as young children and nonhuman animals, are considered exempt. As such, they warrant the taking of an objective stance. Despite their assumed exclusion from moral responsibility practices, moral patients, like those mentioned, are still considered wholly distinct from other exempted entities, such as chairs or cars. This is because the former have intrinsic worth and can be wronged, while the latter are excluded from the moral community altogether. As such, the objective stance is normally assumed to be compatible with accounting for moral patients. However, we do seem to have reason to question the suitability of a wholly objective stance to moral patients.

There is more than one sense that one can participate in moral responsibility practices, and subsequently more than one sense in which one can be recognized or exempted. Acknowledging this distinction is, I believe, key to understanding why a wholly objective stance is detrimental to the stance-taker's sensitivity and responsiveness to moral considerations. According to a popular development of the practice-focused approach, (some) moral responsibility practices are essentially communicative. Engaging with others in these practices thus means taking on roles similar to those in a conversation. When we take the participant stance to someone, we may therefore do so in one, or two, senses. We may see them as someone who is eligible as a recipient of moral reactions, such as blame or ascriptions of moral responsibility. If so, we see them as a moral defendant. But we may also see them as a source or maker of moral reactions, a moral claimant.

I claim that recognizing or exempting someone as a moral claimant disposes the stance-taker very differently toward the moral patient. This is because these perspectives involve distinct other-regarding perspectives. When we take the claimant-directed participant stance to someone, we relate to them as an actual or prospective source or maker of moral claims and demands. In other words, we see them as a potential *you*. They have their own unique evaluative perspective, and may therefore morally appraise and react to us, and others. When we exempt someone as moral claimant, we can only relate to them as someone to regard from outside of the boundaries of a moral exchange. As such, we see them as an object of benevolent concern, as a being with a welfare or interests. I call this the ward perspective.

In this way, including or exempting as moral claimant involve distinct other-regarding perspectives. While I believe that both perspectives are valuable, and complement each other, we seem to have normative reasons to avoid taking a wholly objective stance to moral patients. This is because adding the claimant-directed participant stance to the way we see moral patients disposes us to perceive a wider range of morally relevant facts and considerations. Conversely, wholly exempting someone as moral claimant runs the risk of obscuring any facts or considerations available from the claimant-directed participant stance. We should therefore aim to add the add the claimant perspective to the way we see and relate to all moral patients, and refrain from a pure ward perspective.

There are, I claim, both historical and current examples supporting this moral psychological thesis. Prejudiced thinking and oppressive practices appear to be correlated with views and arrangements where the moral patient in question is exempted as a source or maker of moral claims and demands. Slavery,

institutionalization, and overt paternalism are just a few examples of practices that have been, and, to some extents still are, defended on considerations derived from a primarily wardship basis. Conversely, the value of second-person address is a recurrent theme in recent works exploring systemic bias and discrimination.

A general claimant-directed participant stance can be implemented despite the fact that large populations of moral patients may not meet the requirements for standard moral claimant participation. This is because the claimant perspective, and its disposing effects, are elicited on the mere prospect of being morally addressed. Prospective moral address is, in fact, something we appear to utilize on a day-to-day basis. We often anticipate or imagine what someone may think, feel, and say about the decisions and actions we, and others, make or are considering.

One can say that adopting the claimant perspective is a fundamental part of ordinary human moral deliberation and decision-making. We consult imaginative or real others by engaging with them in imaginative or prospective moral exchanges to gain access to a wider range of potentially relevant facts and considerations. Moreover, the claimant perspective is necessary for the possibility of uptake of actual moral address. Arranging environments to promote the taking of a general claimant-directed participant stance to all moral patients can therefore serve as an avenue for environmentally or socially mediated moral enhancement.

References

Aaltola, E. (2008). Personhood and animals. *Environmental Ethics*, *30*(2), 175-193.

Aaltola, E. (2014). Affective empathy as core moral agency: Psychopathy, autism and reason revisited. *Philosophical Explorations*, *17*(1), 76-92.

Aarskog, N. K., Hunskår, I., & Bruvik, F. (2019). Animal-assisted interventions with dogs and robotic animals for residents with dementia in nursing homes: A systematic review. *Physical & Occupational Therapy in Geriatrics*, *37*(2), 77-93.

Allen, C. & Bekoff, M. (1997). *Species of Mind: The Philosophy and Biology of Cognitive Ethology*. MIT Press.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics?. IEEE *Intelligent Systems*, *21*(4), 12-17. 10.1109/MIS.2006.83

Alm, D. (2023). The Animal Rights Debate Reconsidered. In A. G. Garcia, M. Gunnemyr, & J. Werkmäster (Eds.), *Value, Morality & Social Reality* (pp. 13-23). Lund University. https://doi.org/10.37852/oblu.189.c510

Alston, W. P. (1988). The Deontological Conception of Epistemic Justification. *Philosophical Perspectives*, *2,* 257–299. https://doi.org/10.2307/2214077

ALURES (version date 2020)- European Commission – Animal Use Reporting-Eu System Eu Statistics Database on the Use of Animals for Scientific Purposes under Directive 2010/63/Eu. Available online: https://webgate.ec.europa.eu/envdataportal/content/alures/section1_number-of-animals.html accessed on 1 August 2023

Amaya, S., Doris, J. M. (2015). No Excuses: Performance Mistakes in Morality. In J. Clausen & N. Levy (Eds.), *Handbook of Neuroethics*. Springer, Dordrecht. https://doi-org.ezproxy.ub.gu.se/10.1007/978-94-007-4707-4_120

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge University Press.

Andersson Cederholm, E., Björck, A., Jennbert, K., & Lönngren, A. S. (2014). *Exploring the animal turn: Human-animal relations in science, society and culture.* Pufendorf Institute for Advanced Scholars, Lund University.

Andrews, K. (2020a). *The Animal Mind: An Introduction to the Philosophy of Animal Cognition*. Routledge.

Andrews, K. (2020b). Naïve normativity: The social foundation of moral cognition. *Journal of the American Philosophical Association*, 6(1), 36-56.

Andrews, K. & Gruen, L. (2014). Empathy in Other Apes. In H. Maibon (Ed.), *Empathy and Morality* (pp 193–209). Oxford University Press.

Andrews, K. & Monsó, S. (2021). Animal Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition). Stanford University. https://plato.stanford.edu/archives/spr2021/entries/cognition-animal/ accessed 20 Nov 2023.

Argetsinger, H., & Vargas, M. (2022). What's the Relationship between the Theory and Practice of Moral Responsibility. *HUMANA. MENTE Journal of Philosophical* Studies, *15*(42), 29-62.

Ariely, D. (2009). *Predictably irrational: The hidden forces that shape our decisions.* Harper Collins.

Ariely, D. (2012). *The (honest) truth about dishonesty: How we lie to everyone–especially ourselves.* Harper Collins.

Aristotle. (2002). *Nicomachean Ethics*, (C. J. Rowe, Trans., S. Broadie, Ed.). Oxford University Press. (Original work published 350 B.C.E.)

Arkin, R. C. (2009). Ethical robots in warfare. *IEEE Technology and Society Magazine*, *28*(1), 30-33.

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.

Arpaly, N. (2015). Huckleberry Finn Revisited: Inverse Akrasia and Moral Ignorance. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The Nature of Moral Responsibility: New Essays* (online edn). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199998074.003.0007, accessed 2 Aug. 2023.

Asimov, I. (1942). Runaround. *Astounding science fiction, 29*(1), 94-103.

Austin, R. (2011). *Unmanned aircraft systems: UAVS design, development and deployment*. John Wiley & Sons.

Avramides, A. (2020). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition). Stanford University. https://plato.stanford.edu/archives/win2020/entries/other-minds/, accessed 20 Nov. 2023.

Ayala, F. J. (2010). The difference of being human: Morality. *Proceedings of the National Academy of Sciences*, 107(supplement_2), 9015-9022.

Ayer, A. J. (2013). Freedom and Necessity. In Shafer-Landau, R. (Ed) pp. 16-21. *Ethical theory: An anthology* (2nd ed., Blackwell philosophy anthologies, 17). (Reprinted from *Philosophical Essays*, pp. 271-284, by A. J. Ayer, 1954, Macmillan)

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences,* 201711842. http://www.pnas.org/content/115/25/6506. Accessed 20 Nov. 2023.

Bargh, J.A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist 54*(7), 462-79.

Bargh, J.A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin 126*(6), 925-45.

Bargh, J. A., Schwader, K. L., Hailey, S. E., Dyer, R. L., & Boothby, E. J. (2012). Automaticity in social-cognitive processes. *Trends in cognitive sciences, 16*(12), 593-605.

Batson, C. D. (1994). Why act for the public good? Four answers. *Personality and Social Psychology Bulletin, 20*, 603–610.

Batson, C. D., Ahmad, N., & Stocks, E. L. (2011). Four forms of prosocial motivation: Egoism, altruism, collectivism, and principlism. In D. Dunning (Ed.), *Social motivation* (pp. 103-126). Psychology Press.

Beauchamp, T. L. (1999). Hume on the nonhuman animal. *The Journal of Medicine and Philosophy*, *24*(4), 322-335.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1-3), 7-15.

Behdadi, D. (2019, February) *Conscious AI: A Moral Dilemma*. [video] TEDx Conference Göteborg 2019 – *Disrupting Status Quo*. https://www.ted.com/talks/dorna_behdadi_conscious_ai_a_moral_dilemma , accessed 8 Nov. 2023

Bekoff, M., & Pierce, J. (2009). *Wild justice: The moral lives of animals*: University of Chicago Press.

Benn, S. (1967). Egalitarianism and Equal Consideration of Interests. In J. R. Pennock & J. Chapman (Eds.), *Nomos IX: Equality* (pp. 61–78). Atherton Press.

Bentham, J. (2017). *An Introduction to the Principles of Morals and Legislation* (J. Bennett, Ed.) Retrieved from: www.earlymoderntexts.com. (Original work published 1789)

Benz-Schwarzburg, J., & Wrage, B. (2023). Caring animals and the ways we wrong them. *Biology & Philosophy, 38*(4), 25. https://doi-org.ezproxy.ub.gu.se/10.1007/s10539-023-09913-1

Beran, M. J., Parrish, A. E., Perdue, B. M., & Washburn, D. A. (2014). Comparative cognition: Past, present, and future. *International Journal of Comparative Psychology. 27*(1), 3-30.

Bergson, H. (1910). *Time and free will,* (F. L. Pogson, Trans.). Allen and Unwin. (Original work published 1889)

Berkeley, G. (1998). *A Treatics Concerning the Principles of Human Knowledge.* (Jonathan Dancy, Ed.), Oxford University Press. (Original work published 1710)

Bermúdez, J. L. (2003). The domain of folk psychology. In A. O'Hear (Ed.), *Minds and Persons* (pp. 1–29). Cambridge University Press.

Bernstein, I. S. (2000). The law of parsimony prevails. Missing premises allow any conclusion. *Journal of Consciousness Studies, 7*(1-2), 31-34.

Best, S. (2009). The rise of critical animal studies: Putting theory into action and animal liberation into higher education. *Journal for Critical Animal Studies, 7*(1), 9-52.

Best, S., Nocella, A. J., Kahn, R., Gigliotti, C., & Kemmerer, L. (2007). Introducing critical animal studies. *Journal for Critical Animal Studies, 5*(1), 4-5.

Bickle, J., (2020). Multiple Realizability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition). Stanford University. https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability/

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21-34.

Birch, J., Burn, C., Schnell, A., Browning, H., Crump, A. (2021). *Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans.* LSE Consulting. WBI Studies Repository. https://www.wellbeingintlstudiesrepository.org/cgi/viewcontent.cgi?article=1001&context=af_gen

Birhane, A., & van Dijk, J. (2020). Robot rights? Let's talk about human welfare instead. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 207-213).

Björnsson, G. (2017). Explaining (Away) the Epistemic Condition on Moral Responsibility. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 146–162). Oxford University Press.

Blattner, C. E., Coulter, K., & Kymlicka, W. (Eds.). (2019). *Animal labour: A new frontier of interspecies justice?*. Oxford University Press.

Block, N. (1995). How many concepts of consciousness?. *Behavioral and Brain Sciences, 18*(2), 272-287.

Boddington, P. (2021). AI and moral thinking: how can we live well with machines to enhance our moral agency?. *AI and Ethics, 1*, 109-111.

Borchert, R., & Dewey, A. R. (2023). In praise of animals. *Biology & Philosophy, 38*(4), 24.

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In R. V. Yampolskiy (Ed.) *Artificial intelligence safety and security* (pp. 57-69). Chapman and Hall/CRC.

Brandenburg, D. (2018). The nurturing stance: Making sense of responsibility without blame. *Pacific Philosophical Quarterly, 99*, 5-22.

Brandenburg, D. (2019). Inadequate agency and appropriate anger. *Ethical Theory and Moral Practice, 22*(1), 169-185.

Brink, D. O. &. Nelkin, D. K. (2013). Fairness and the Architecture of Responsibility. In D. Shoemaker & N. A. Tognazzini (Eds.), *Oxford studies in agency and responsibility* (Vol. 1, pp. 284–314). Oxford University Press. doi:10.1093/acprof:oso/9780199694853.003.0013

Brosnan, S. F., & de Waal, F. B. (2002). A proximate perspective on reciprocal altruism. *Human Nature*, *13*, 129-152.

Brouwer, D. D. Z. (2018). *Neither Wild nor Domesticated: Positioning Liminal Animals through Labour Rights* [Doctoral dissertation, Queen's University, Canada]. QSpace: Queen's Scholarship & Digital Collections. https://qspace.library.queensu.ca/server/api/core/bitstreams/4443b00c-fa3d-4ae5-87a8-186ac7625cd7/content

Browning, H., & Birch, J. (2022). Animal sentience. *Philosophy compass, 17*(5), e12822.

Buchanan, J. (2000). The Limits of Liberty: Between Anarchy and Leviathan. In. J. Buchanan (Ed.), *The Collected Works of James M. Buchanan* (Vol. 7). Liberty Fund, Inc. (Original work published 1975)

Buckner, C. (2013). Morgan's Canon, meet Hume's Dictum: avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy, 28*, 853-871.

Buckner, C. (2021). Black boxes, or unflattering mirrors? Comparative bias in the science of machine behavior. *The British Journal for the Philosophy of Science. 74*(3), 681-712.

Burgis, J. L. (2018). *Making Covenants with Brute Beasts: Making Room for Nonhuman Animals in a Contractualist Framework* [Doctoral dissertation, University of Miami]. Available online: https://scholarship.miami.edu/esploro/outputs/doctoral/Making-Covenants-with-Brute-Beasts-Making/991031447234502976/filesAndLinks?index=0

Burke, K. A., Oron-Gilad, T., Conway, G., & Hancock, P. A. (2007). Friend/foe identification and shooting performance: Effects of prior task loading and time pressure. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 51, No. 4, pp. 156-160). SAGE Publications.

Burroughs, M. D. (2020). Navigating the penumbra: Children and moral responsibility. *The Southern Journal of Philosophy, 58*(1), 77-101.

Calhoun, C. (1989). Responsibility and reproach. *Ethics, 99*(2), 389-406.

Calhoun, C. (2015). But What About the Animals?. In M. Timmons. & R. N. Johnson (Eds.), *Reason, Value, and Respect: Kantian Themes from the Philosophy of Thomas E. Hill, Jr.* (online edn). Oxford University Press. https://doi-

org.ezproxy.ub.gu.se/10.1093/acprof:oso/9780199699575.003.0011, accessed 10 Aug. 2023.

Campbell, A. L., Naik, R. R., Sowards, L., & Stone, M. O. (2002). Biological infrared imaging and sensing. *Micron, 33*(2), 211-225.

Campbell, J. (2017). PF Strawson's free will naturalism. *International Journal for the Study of Skepticism, 7*(1), 26-52.

Carbonell, V. (2019). Social Constraints On Moral Address. *Philosophy and Phenomenological Research, 98*(1), 167-189.

Carlson, L. (2009). *The faces of intellectual disability: Philosophical reflections.* Indiana University Press.

Carruthers, P. (1992). *The Animals Issue.* Cambridge University Press.

Carruthers, P. (2019). *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford University Press. doi:10.1093/oso/9780198843702.001.0001

Charles, N., & Davies, C. A. (2008). My family and other animals: Pets as kin. *Sociological Research Online, 13*(5), 13-26.

Charles, N., Fox, R., Miele, M., & Smith, H. (2022). Dogs at Work: Gendered Organizational Cultures and Dog-Human Partnerships. In L. Tallberg & L. Hamilton (Eds.), *The Oxford Handbook of Animal Organization Studies*, (online edn). Oxford Academic. https://doi-org.ezproxy.ub.gu.se/10.1093/oxfordhb/9780192848185.013.29, accessed 1 Aug. 2023.

Chisholm, R. M. (1966). Freedom and Action. In K. Lehrer (Ed.), *Freedom and Determinism* (pp. 11-44). Random House.

Ciurria, M. (2014). Moral Responsibility: Justifying Strawson and the Excuse of Peculiarly Unfortunate Formative Circumstances. *Ethical Theory and Moral Practice, 17*, 545-557.

Ciurria, M. (2023). Responsibility's Double Binds: The Reactive Attitudes in Conditions of Oppression. *Journal of Applied Philosophy*, *40*(1), 35-48.

Clark, S. R. L. (1984). *The nature of the beast: Are animals moral?*. Oxford University Press.

Clark, S. R. L. (1985). Good Dogs and Other Animals. In. P. Singer (Ed.), *In Defense of Animals.* Basil Blackwell.

Clarke, R., Capes, J. & Swenson, P. (2021). Incompatibilist (Nondeterministic) Theories of Free Will. In E. N. Zalta, (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition). Stanford University. https://plato.stanford.edu/archives/fall2021/entries/incompatibilism-theories/

Claudy, M. C., Aquino, K., & Graso, M. (2022). Artificial intelligence can't be charmed: the effects of impartiality on laypeople's algorithmic preferences. *Frontiers in Psychology, 13*, 898027.

Clement, G. (2013). Animals and moral agency: The recent debate and its implications. *Journal of Animal Ethics, 3*(1), 1-14.

Cochrane, A., Garner, R., & O'Sullivan, S. (2018). Animal ethics and the political. *Critical Review of International Social and Political Philosophy, 21*(2), 261-277.

Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society, 24*(2), 181–189.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241.

Coeckelbergh, M. (2020a). *AI ethics*. MIT Press.

Coeckelbergh, M. (2020b). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics, 26*(4), 2051-2068.

Cova, F. (2013). Two kinds of moral competence: Moral agent, moral judge. In B. Musschenga & A. van Harskamp (Eds.), *What Makes Us Moral? On the capacities and conditions for being moral* (pp. 117-130). Springer.

Coveney, P. (1982). The Image of the Child. In C. Jenks (Ed.), *The Sociology of Childhood: Essential Readings* (pp. 42–47). Batsford Academic and Educational Ltd.

Cudd, A. & Eftekhari, S. (2021). Contractarianism. In. E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). Stanford University. https://plato.stanford.edu/archives/win2021/entries/contractarianism/

Damasio, A. (1994). *Descartes' Error: Emotion, reason and the human brain*. GP Putnam's Sons.

Damasio, A. R., Tranel, D., & Damasio, H. C. (1991). Somatic markers and the guidance of behavior: theory and preliminary testing. In H. S. Levin, H. M. Eisenberg & L. B. Benton (Eds.), *Frontal Lobe Function and Dysfunction* (pp. 217-229). Oxford University Press.

Danaher, J. (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and engineering ethics, 26*(4), 2023-2049.

Danón, L. (2019). Animal Normativity. *Phenomenology and Mind, (17*), 176–187. https://doi.org/10.13128/pam-8035

Darley, J. M., & Batson, C. D. (1973). "From Jerusalem to Jericho": A study of situational and dispositional variables in helping behavior. *Journal of personality and social psychology, 27*(1),100.

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. In R. Calo, A. Froomkin, & I. Kerr (Eds.), *Robot law* (pp. 213–232). Edward Elgar Publishing.

Darwall, S. (1977). Two Kinds of Respect. *Ethics 88* (1): 36–49.

Darwall S. (2006) *The Second-Person Standpoint*. Harvard University Press.

Dashper, K. (2016). *Human-animal relationships in equestrian sport and leisure*. Taylor & Francis.

Davidson, D. (1982). Rational animals, *Dialectica, 36*: 317–327.

Davidson, D. (1984). *Inquiries into truth and interpretation*. Clarendon.

Davis, J. (2022). *Adultification Bias Within Child Protection and Safeguarding*, HM Inspectorate of Probation; available online at https://www.justiceinspectorates.gov.uk/hmiprobation/wp-content/ uploads/sites/5/2022/06/Academic-Insights-Adultification-bias-within-child-protection-andsafeguarding.pdf

Davis, J. & Marsh, N. (2020). Boys to men: the cost of 'adultification' in safeguarding responses to Black boys. *Critical and Radical Social Work, 8*(2), 255-259.

Decety, J., & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in cognitive sciences, 18*(7), 337-339.

Decety, J., Bartal, I. B. A., Uzefovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behaviour: highly conserved neurobehavioural mechanisms across species. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1686), 20150077.

DeGrazia, D. (1996). *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge University Press.

DeGrazia, D. (2008). Moral status as a matter of degree?. *The Southern Journal of Philosophy, 46*(2), 181-198.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science (American Association for the Advancement of Science), 358*(6362), 486-492.

Delon, N. (in press). Letting animals off the hook. *Journal of Ethics and Social Philosophy*

DeMello, M. (2021). *Animals and society: An introduction to human-animal studies*. Columbia University Press.

De Mesel, B. (2022). Being and holding responsible: Reconciling the disputants through a meaning-based Strawsonian account. *Philosophical Studies*, *179*(6), 1893-1913.

Denis, L. (2000). Kant's conception of duties regarding animals: reconstruction and reconsideration. *History of Philosophy Quarterly, 17*(4), 405-423.

Dennett, D. (1984). *Elbow Room.* MIT Press

de Waal, F. (1996). *Good natured: The origins of right and wrong in humans and others.* Harvard University Press.

de Waal, F. (2006). *Primates and philosophers: How morality evolved.* Princeton University Press.

de Waal, F. (2010). *The age of empathy: Nature's lessons for a kinder society.* Crown.

de Waal, F. (2014). Natural normativity: The 'is' and 'ought'of animal behavior. *Behaviour, 151*(2-3), 185-204.

Dick, P. K. (1968). *Do androids dream of electric sheep?*. Rapp & Whiting.

Dietrich, E. (2001). Homo sapiens 2.0: Why we should build the better robots of our nature. *Journal of Experimental & Theoretical Artificial Intelligence, 13*(4), 323-328.

Dixon, B. (1995). Response: Evil and the moral agency of animals. *Between the Species, 11*(1–2), 38–40.

Dixon, B. (2008). *Animals, emotion & morality: Marking the boundary.* Amherst, NY: Prometheus Books.

Donaldson, S., & Kymlicka, W. (2011). *Zoopolis: A political theory of animal rights.* Oxford University Press.

Donaldson, S., & Kymlicka, W. (2016). Comment: Between wild and domesticated: Rethinking categories and boundaries in response to animal agency. In B. Bovenkerk & J. Keulartz (Eds.), *Animal ethics in the age of humans: Blurring boundaries in human-animal relationships* (pp. 225-239). Springer Cham. https://doi.org/10.1007/978-3-319-44206-8_14

Döring, N., Mohseni, M.R. & Walter, R. (2020). Design, Use, and Effects of Sex Dolls and Sex Robots: *Scoping Review J Med Internet Res 22*(7): e18551

Doris, J. M. (2002). *Lack of character: Personality and moral behavior.* Cambridge University Press.

Doris, J. M. (2015). *Talking to our selves: Reflection, ignorance, and agency.* Oxford University Press.

Duncan, I. J. (2006). The changing concept of animal sentience. *Applied Animal Behaviour Science, 100*(1–2), 11–19. https:// doi.org/10.1016/j.applanim.2006.04.011

Dworkin, R. (1993). *Life's Dominion: An Argument about Abortion, Euthanasia, and Individual Freedom.* Vintage Books.

Ekstrom, L. (1993). A Coherence Theory of Autonomy, *Philosophy and Phenomenological Research, 53*: 599–616.

Ekstrom, L. (2000). *Free Will: A Philosophical Study*. Westview Press.

Ekstrom, L. (2003). Free Will, Chance, and Mystery, *Philosophical Studies, 113*: 153–80.

El-Alti, L. (2023). Shared Decision Making in Psychiatry: Dissolving the Responsibility Problem. *Health Care Analysis, 31*(2), 65-80.

Elder, A. (2017). Robot Friends for Autistic Children: Monopoly Money or Counterfeit Currency?, In P. Lin, K. Abney & R. Jenkins (Eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (online edn.). Oxford Academic. https://doi.org/10.1093/oso/9780190652951.003.0008, accessed 1 Aug. 2023.

Epley, N., & Waytz, A. (2010). Mind perception. In S. Fiske, D. Gilbert & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. 2, pp. 498-541). Wiley.

Ferrin, A. (2019). Nonhuman animals are morally responsible. *American Philosophical Quarterly, 56*(2), 135-154.

Fields, L. (1994). Moral beliefs and blameworthiness: Introduction. *Philosophy, 69*(270), 397-415.

Fine, C., & Kennett, J. (2004). Mental impairment, moral understanding and criminal responsibility: Psychopathy and the purposes of punishment. *International Journal of Law and Psychiatry, 27*(5), 425-443.

Fischer, J. M., & Ravizza, M. (1998*). Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fitzpatrick, S. (2017). Animal morality: What is the debate about?. *Biology & Philosophy, 32*(6), 1151-1183.

FitzPatrick, W. J. (2008). Moral responsibility and normative ignorance: Answering a new skeptical challenge. *Ethics, 118*(4), 589-613.

Flack, J. C., & de Waal, F. B. (2000). 'Any animal whatever'. Darwinian building blocks of morality in monkeys and apes. *Journal of Consciousness Studies, 7*(1-2), 1-29.

Floridi, L. (2023). AI as Agency without Intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology, 36*(1), 15

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines, 14*, 349-379.

Fox, R. (2006). Animal behaviours, post-human lives: Everyday negotiations of the animal–human divide in pet-keeping. *Social & Cultural Geography, 7*(4), 525-537.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy, 66*(23), 829-839.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy, 68*(1), 5-20.

Franklin, S. (2014). History, motivations, and core themes. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 15-33). Cambridge University Press. doi:10.1017/CBO9781139046855.003

Fricker, M. (2016). What's the point of blame? A paradigm based explanation. *Noûs 50*/1:165–183. https://doi.org/10.1111/nous.12067

Fritz, A., Brandt, W., Gimpel, H., & Bayer, S. (2020). Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). *De Ethica, 6*(1), 3-22.

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY, 35,* 795-809.

Gauthier, D. (1986). *Morals By Agreement*. Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Ginet, C. (1966). Might We Have No Choice?, In K. Lehrer (Ed.), *Freedom and Determinism* (pp. 87–104). Random House.

Ginet, C. (1989). Reasons Explanations of Action: An Incompatibilist Account, *Philosophical Perspectives, 3*: 17–46.

Ginet, C. (1990). *On Action*. Cambridge University Press.

Ginet, C. (1997). Freedom, responsibility, and agency. *The Journal of Ethics, 1*, 85-98.

Gips., J. (1995). Towards the Ethical Robot. In K. Ford, C. Glymour & P. Hayes, (Eds.) *Android Epistemology* (pp. 243-252). MIT Press.

Goff, P. A., Jackson, M. C., Di Leone, B. A. L., Culotta, C. M., & DiTomasso, N. A. (2014). The essence of innocence: consequences of dehumanizing Black children. *Journal of personality and social psychology*, *106*(4), 526-45.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological inquiry*, *23*(2), 101-124.

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.

Greenspan, P.S. (2003) Responsible psychopaths, *Philosophical Psychology, 16*:3, 417-429. 10.1080/0951508032000121797

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California law review, 94*(4), 945-967.

Griffith, M. (2010). Why agent-caused actions are not lucky. *American Philosophical Quarterly, 47*(1), 43-56.

Gruen, L. (2017). Conscious Animals and the Value of Experience, In S. M. Gardiner & A. Thompson (Eds.), *The Oxford Handbook of Environmental Ethics* (p. 91-100). Oxford Academic. https://doi-org.ezproxy.ub.gu.se/10.1093/oxfordhb/9780199941339.013.9, accessed 9 Aug. 2023.

Gruen, L. (2021). The Moral Status of Animals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/archives/sum2021/entries/moral-animal/

Gunkel, D. J. (2018a). The relational turn: third wave HCI and phenomenology. In M. Filimowicz & V. Tzankova (Eds.), *New Directions in Third Wave Human-Computer Interaction* (Vol. 1-Technologies, pp. 11-24). Springer Cham.

Gunkel, D. J. (2018b). *Robot Rights*. MIT Press.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review, 108*(4), 814.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Haji, I. (1997). An Epistemic Dimension of Blameworthiness, *Philosophy and Phenomenological Research, 57*(3): 523–544. doi:10.2307/2953747

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259-275.

Haksar, V. (1998). Moral agents. In E. Craig (Ed.), *Routledge Encyclopedia of Philosophy (Vol. 6*, pp. 499-504). Routledge Press.

Hallamaa, J., & Kalliokoski, T. (2020). How AI systems challenge the conditions of moral agency?. *International Conference on Human-Computer Interaction* (pp. 54-64). Cham Springer International Publishing.

Hallgren, I. (2012). Seeing agents when we need to, attributing experience when we feel like it. *Review of Philosophy and Psychology*, *3*, 369-382.

Harman, E. (2011). Does moral ignorance exculpate?. *Ratio, 24*(4), 443-468.

Hartvigsson, T. & Munthe, C. (2018) *Responsibilities in change: modelling parental authority and children's autonomy*. Unpublished manuscript, ResearchGate. https://www.researchgate.net/publication/316968012_Responsibilities_in_change_modelling_parental_authority_and_children's_autonomy accessed Aug. 7 2023.

Heilig, M., MacKillop, J., Martinez, D., Rehm, J., Leggio, L., & Vanderschuren, L. J. (2021). Addiction as a brain disease revised: why it still matters, and the need for consilience. *Neuropsychopharmacology, 46*(10), 1715-1723.

Heyes, C. (2023). Rethinking Norm Psychology. *Perspectives on Psychological Science*, *0*(0). https://doi.org/10.1177/17456916221112075

Hickman, D.L., Johnson, J., Vemulapalli, T.H., Crisler, J.R. & Shepherd, R. (2017). Commonly Used Animal Models. In M. A. Suckow & K. L. Stewart (Eds.), *Principles of Animal Research for Graduate and Undergraduate Students* (pp. 117–75). Academic Press. 10.1016/B978-0-12-802151-4.00007-4. Epub 2016 Nov 25. PMCID: PMC7150119.

Hieronymi, P. (2008). Responsibility for believing. *Synthese, 161*, 357-373.

Hieronymi, P. (2020). *Freedom, resentment, and the metaphysics of morals*. Princeton University Press.

Hill, T. E. (1997). Respect for humanity. In G. Peterson (Ed.), *Tanner Lectures on Human Values,* (18, 1-76). University of Utah Press. https://tannerlectures.utah.edu/_resources/documents/a-to-z/h/Hill97.pdf

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?. *Ethics and Information Technology, 11*, 19-29.

Hobbes, T. (1997). *Leviathan* (R.E. Flatman & D. Johnston, Eds.). W.W. Norton & Co. (Original work published 1651)

Hobbes, T. (1999). Of Liberty and Necessity. In V. Chappell (Ed.), *Hobbes and Bramhall on Liberty and Necessity* (pp. 15.42). Cambridge University Press. (Original work published 1654)

Holroyd, J. (2018) Two ways of socializing moral responsibility: Circumstantialism versus scaffolded- responsiveness. In K. Hutchison, C. Mackenzie & M. Oshana (Eds.) *Social Dimensions of Moral Responsibility* (pp. 137-162). Oxford University Press.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian journal of philosophy, 72*(1), 63-76.

Horowitz, A. (2016). *Being a dog: Following the dog into a world of smell*. Simon and Schuster.

Hortensius, R., & Cross, E. S. (2018). From automata to animate beings: the scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences, 1426*(1), 93-110.

Hudson, S. D. (1980). The nature of respect. *Social Theory and Practice, 6*(1), 69-90.

Hume, D. (1875). A Dissertation on the Passions. In T. H. Green & T. H. Grose (Eds.), *Essays: Moral, Political and Literary* 2 vols. (Vol. 2, pp. 137-166). Longmans, Green. (Original work published 1757)

Hume, D. (1975). *Enquiries concerning Human Understanding and concerning the Principles of Morals (*L. A. Selby-Bigge & P. H. Nidditch, Eds.). Clarendon Press. (Original work published 1777)

Hume, D. (1978), *A Treatise of Human Nature* (L. A. Selby-Bigge & P. H. Nidditch, Eds.). Clarendon Press. (Original work published 1739–1740)

Husak, D. (2011). Negligence, Belief, Blame, and Criminal Liability: The Special Case of Forgetting. *Criminal Law and Philosophy, 5*(2): 199–218. doi:10.1007/s11572-011-9115-z

Hutchison, K. (2018). Moral Responsibility, Respect, and Social Identity. In K. Hutchison, C. Mackenzie & M. Oshana (Eds.), *Social Dimensions of Moral Responsibility* (online edn., pp. 206-230). Oxford Academic. https://doi-org.ezproxy.ub.gu.se/10.1093/oso/9780190609610.003.0009, accessed 9 Aug. 2023.

Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of personality and social psychology, 21*(3), 384.

J. H. Moor. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems, vol. 21*, no. 4, 18-21. 10.1109/MIS.2006.80.

Jamieson, D., & Bekoff, M. (1992). On aims and methods of cognitive ethology. *PSA: proceedings of the biennial meeting of the Philosophy of Science Association (Vol. 1992*, No. 2, pp. 110-124). Philosophy of Science Association.

Jaworska, A. (2007). Caring and internality. *Philosophy and Phenomenological Research, 74*(3), 529-568.

Jaworska, A. & Tannenbaum, J. (2023). The Grounds of Moral Status. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). Stanford University. https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/

Jefferson, A. (2019). Instrumentalism about moral responsibility revisited. *The Philosophical Quarterly, 69*(276), 555-573.

Jennings, B. (2009). Agency and moral relationship in dementia. *Metaphilosophy, 40*(3-4), 425-437.

Jensen, M. E., Moss, C. F., & Surlykke, A. (2005). Echolocating bats can use acoustic landmarks for spatial orientation. *Journal of experimental biology, 208*(23), 4399-4410.

Jensen, P. (Ed.). (2017). *The ethology of domestic animals: an introductory text*. Cabi.

Jeppsson, S. (2021). Psychosis and Intelligibility. *Philosophy, Psychiatry, & Psychology 28*(3), 233-249. https://doi.org/10.1353/ppp.2021.0036.

Jeppsson, S. (2022a). My strategies for dealing with radical psychotic doubt: A schizo-something philosopher's tale. *Schizophrenia Bulletin, 49*(5), 1097-1098.

Jeppsson, S. (2022b) Accountability, Answerability, and Attributability: On Different Kinds of Moral Responsibility. In D. K. Nelkin, & D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (pp. 73-88). Oxford University Press. https://doi-org.ezproxy.ub.gu.se/10.1093/oxfordhb/9780190679309.013.21, accessed 8 Nov. 2023.

Jeppsson, S. (2023). A Wide-Enough Range of 'Test Environments' for Psychiatric Disabilities. *Royal Institute of Philosophy Supplements*, *94*, 39-53.

Jeppsson, S. (in press). Ciurria and Strawson: how deep is the divide. *Syndicate Philosophy. Advance online publication*. https://philpapers.org/archive/JEPCAS.pdf

Johansson, L. (2010). The functional morality of robots. *International Journal of Technoethics, 1*(4), 65–73.

Johnsen, S., & Lohmann, K. J. (2005). The physics and neurobiology of magnetoreception. *Nature reviews neuroscience, 6*(9), 703-712.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8,* 195-204.

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology, 7*(2), 99–107.

Johnson, D. G., & Verdicchio, M. (2018). Why robots should not be treated like animals. *Ethics and Information Technology, 20*, 291-301.

Jones, K. (2003). Emotion, Weakness of Will, and the Normative Conception of Agency. *Royal Institute of Philosophy Supplements, 52*, 181-200.

Jones, R. C. (2013). Science, sentience, and animal welfare. *Biology and Philosophy, 28*, 1–30. https://doi.org/10.1007/ s10539-012-9351-1

Joyce, R. A. (2006). *The evolution of morality*. MIT Press.

Kagan, J. (2000). Human morality is distinctive. *Journal of Consciousness Studies, 7*(1-2), 46-48.

Kahneman, D. (2011). *Thinking fast and slow*. New York: Farrar, Straus, and Giroux.

Kain, P. (2009). Kant's Defense of Human Moral Status, *Journal of the History of Philosophy, 47*, 59–102.

Kalof, L., & Whitley, C. T. (2021). Animals in environmental sociology. In B. Caniglia, A. Jorgenson, S. Malin, L. Peek & D. Pellow (Eds.), *International Handbook of Environmental Sociology* (pp. 289-313). Springer, Cham. https://doi-org.ezproxy.ub.gu.se/10.1007/978-3-030-77712-8_14

Kane, R. (1996). *The Significance of Free Will*. Oxford University Press.

Kant, I. (1963) *Lectures on Ethics* (H. L. Infield, Trans.). Harper and Row. (Original work published 1780–1)

Kant, I. (1987). *Critique of pure reason* (P. Guyer & A. Wood, Trans.). Cambridge University Press. (Original work published 1781/1787)

Kant, I. (1996). Groundwork of The metaphysics of morals. In M. Gregor (Ed.), P*ractical Philosophy (*The Cambridge Edition of the Works of Immanuel Kant, pp. 37-108). Cambridge University Press. (Original work published 1785)

Kant, I. (1996). Critique of practical reason. In M. Gregor (Ed.), *Practical Philosophy* (The Cambridge Edition of the Works of Immanuel Kant, pp. 133-272). Cambridge University Press. (Original work published 1788)

Kant, I. (1996). The metaphysics of morals. In M. Gregor (Ed.), *Practical Philosophy (*The Cambridge Edition of the Works of Immanuel Kant*,* pp. 353-604). Cambridge University Press. (Original work published 1797)

Kant, I. (2010) Anthropology from a Pragmatic Point of View. In R. Louden & G. Zoller (Eds. and Trans.), *Anthropology, History, and Education*, (Cambridge Edition of the Works of Immanuel Kant, pp. 227-429). Cambridge University Press. (Original work published 1798)

Kauppinen, A. (2017). Empathy and moral judgment. In H. L. Maibom (Ed.), *The Routledge Handbook of Philosophy of Empathy* (pp. 215-226). Routledge.

Kauppinen, A. (2022). Moral Sentimentalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/archives/spr2022/entries/moral-sentimentalism/

Kennett, J. (2002). Autism, empathy and moral agency. *The Philosophical Quarterly, 52*(208), 340-357.

Kennett, J. (2006). Do psychopaths really threaten moral rationalism?. *Philosophical Explorations, 9*(1), 69-82.

Kennett, J. (2009). Mental Disorder, Moral Agency, and the Self. In B. Stein (Ed.), *The Oxford Handbook of Bioethics* (pp. 91-113). Oxford University Press.

Kennett, J., & Wolfendale, J. (2019). Self-Control and Moral Security. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility* (Vol. 6, pp. 33-63). Oxford University Press.

Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education. *International Journal of Human–Computer Interaction, 36*(20), 1902-1911.

Kitcher, P. (2006). Ethics and evolution: how to get here from there. In Ober J, Macedo S (Eds.), *Primates and philosophers: how morality evolved* (pp. 120-139). Princeton University Press.

Kitcher, P. (2011). *The ethical project*. Harvard University Press.

Kittay, E. F. (1999). *Love's labor: Essays on women, equality and dependency*. Routledge.

Kittay, E. F. (2009). The personal is philosophical is political: A philosopher and mother of a cognitively disabled person sends notes from the battlefield. *Metaphilosophy, 40*(3-4), 606-627.

Klincewicz, M. (2015). Autonomous weapons systems, the frame problem and computer security. *Journal of Military Ethics*, *14*(2), 162-176.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908-911.

Köhler, S., Roughley, N., & Sauer, H. (2017). Technologically blurred accountability?: Technology, responsibility gaps and the robustness of our everyday conceptual scheme. In C. Ulbert et al. (Eds.), *Moral agency and the politics of responsibility* (pp. 51-68). Routledge.

Korsgaard, C. M. (1996). *The Sources of Normativity*. Cambridge University Press.

Korsgaard, C. M. (2004). Fellow creatures: Kantian ethics and our duties to animals. *Tanner lectures on human values. 24*: 77-110. http://nrs.harvard.edu/urn-3:HUL.InstRepos:3198692, accessed 8 Nov. 2023.

Korsgaard, C. M. (2006). Morality and the distinctiveness of human action. In J. Ober & S. Macedo (Eds.) *Primates and philosophers: how morality evolved* (pp. 98-119). Princeton University Press.

Korsgaard, C. M. (2010). *Reflections on the Evolution of Morality.* Amherst Lecture in Philosophy. The Department of Philosophy at Amherst College. http://nrs.harvard.edu/urn-3:HUL.InstRepos:5141952, accessed 8 Nov. 2023.

Korsgaard, C. M. (2018a). *Fellow creatures: Our obligations to the other animals.* Oxford University Press.

Korsgaard, C. M. (2018b). The claims of animals and the needs of strangers: two cases of imperfect right. *Journal of Practical Ethics, 6*(1).

Krahn, T., & Fenton, A. (2009). Autism, empathy and questions of moral agency. *Journal for the Theory of Social Behaviour, 39*(2), 145-166.

Kymlicka, W., & Donaldson, S. (2014). Animals and the Frontiers of Citizenship. *Oxford Journal of Legal Studies, 34*(2), 201-219.

Kymlicka, W., & Donaldson, S. (2017). Inclusive Citizenship Beyond the Capacity Contract. In A. Shachar et al., (Eds.), *The Oxford Handbook of Citizenship* (online edn., pp. 838-860). Oxford Academic. https://doi-org.ezproxy.ub.gu.se/10.1093/oxfordhb/9780198805854.013.36, accessed 10 Aug. 2023.

Lamb, J. W. (1977). On a Proof of Incompatibilism. *The Philosophical Review, 86*(1), 20–35. doi:10.2307/2184160

Laurence, S. & Margolis, E. (2012). The Scope of the Conceptual. In E. Margolis, R. Samuels, & S. Stich (Eds.), *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 291-317). Oxford University Press.

Levy, N. (2003). Cultural membership and moral responsibility. *The Monist*, *86*(2), 145-163.

Levy, N. (2007). The responsibility of the psychopath revisited. *Philosophy, Psychiatry, & Psychology, 14*(2), 129-138.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford University Press.

Levy, N. (2014). *Consciousness and moral responsibility*. Oxford University Press.

Lewis, S. L. & Maslin, M. A. (2015). Defining the Anthropocene. *Nature*, *519*(7542), 171-180.

Lima, G., Grgić-Hlača, N., & Cha, M. (2021, May). Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (*pp. 1-17).

Lippert-Rasmussen, K. (2003). Identification and responsibility. *Ethical Theory and Moral Practice, 6*, 349-376.

Lorini, G. (2022). Animal Norms: An Investigation of Normativity, *The Non-Human Social World. Law, Culture and the Humanities, 18*(3), 652–673. https://doi.org/10.1177/1743872118800008

Loughnan, S., & Haslam, N. (2007). Animals and androids: Implicit associations between social categories and nonhumans. *Psychological science, 18*(2), 116-121.

Low, P., Panksepp, J., Reiss, D., Edelman, D. B., Van Swinderen, B., & Christof, K. (2012). *The Cambridge Declaration on Consciousness.* http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf, accessed July 31, 2023.

MacDougall-Shackleton, S. A. (2011). The levels of analysis revisited. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1574), 2076-2085.

Mackenzie, C. (2018). Moral responsibility and the social dynamics of power and oppression. In K. Hutchison, C. Mackenzie, & M. Oshana (Eds.), *Social dimensions of moral responsibility* (pp. 59-80). Oxford University Press.

MacLennan, B. (2013). Cruelty to robots? The hard problem of robot suffering. *Proceedings of the 2013 Meeting of the International Association for Computing and Philosophy (IACAP).*

Macnamara, C. (2011). Holding others responsible. *Philosophical Studies, 152*, 81-102.

Macnamara, C. (2013). Taking Demands Out of Blame. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 141-161). Oxford Academic. https://doi.org/10.1093/acprof:oso/9780199860821.003.0008, accessed 6 Nov. 2023.

Macnamara, C. (2015a). Reactive attitudes as communicative entities. *Philosophy and Phenomenological Research, 90*(3), 546-569.

Macnamara, C. (2015b). Blame, communication, and morally responsible agency. In M. Clarke, M. McKenna & A. Smith (Eds.), *The nature of moral responsibility: New essays* (pp. 211-236). Oxford University Press.

Malle, B. F. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 297-352). Academic Press.

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: new answers to an old question. *Journal of personality and social psychology, 93*(4), 491.

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. A. Ferreira, S. J. Sequeira, S. G. Virk, M. Tokhi, E. E. Kadar (Eds.), *Robotics and Well-Being. Intelligent Systems, Control and Automation: Science and Engineering* (Vol 95, pp. 111-133). Springer, Cham. https://doi-org.ezproxy.ub.gu.se/10.1007/978-3-030-12524-0_11*Robotics and well-being,*

Markovits, J. (2010). Acting for the right reasons. *Philosophical Review, 119*(2), 201-242.

Martini, M. C., Buzzell, G. A., & Wiese, E. (2015). Agent appearance modulates mind attribution and social attention in human-robot interaction. In A. Tapus, E. André, J. C. Martin, F. Ferland, M. Ammi (Eds.), *Social Robotics. ICSR 2015. Lecture Notes in Computer Science* (Vol. 9388, pp. 431-439). Springer. https://doi-org.ezproxy.ub.gu.se/10.1007/978-3-319-25554-5_43

Mason, E. (2015). Moral Ignorance and Blameworthiness. *Philosophical Studies, 172*(11): 3037–3057. doi:10.1007/s11098-015-0456-7

Mason, E. (2017). Moral Incapacity and Moral Ignorance. In. R. Peels (Ed.), *Perspectives on Ignorance from Moral and Social Philosophy* (pp. 30-52). Routledge.

Mason, E. (2019). *Ways to be blameworthy: Rightness, wrongness, and responsibility.* Oxford University Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Mayr, E. (1961). Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, *134*(3489), 1501-1506.

McCann, H. J. (1998). *The Works of Agency: On Human Action, Will, and Freedom*. Cornell University Press.

McGeer, V. (2008). Varieties of moral agency: lessons from autism (and psychopathy). In W. Sinnott-Armstrong (Ed.), *Moral psychology. emotion, brain disorders, and developmen*t - *The neuroscience of morality* (Vol. 3, pp. 227-257). MIT Press.

McGeer, V. (2009). The Skill of Perceiving Persons. *Modern Schoolman, 86*(3-4), 289-318.

McGeer, V. (2012). Co-reactive attitudes and the making of moral community. In R. Langdon & C. Mackenzie (Eds.), *Emotions, imagination and moral reasonin*g (pp. 299-326). Psychology Press.

McGeer, V. (2013). Civilizing Blame. In J. D. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 162-188). Oxford University press.

McGeer, V. (2019). Scaffolding agency: A proleptic account of the reactive attitudes. *European Journal of Philosophy, 27*(2), 301-323.

McGeer, V., & Pettit, P. (2015). The Hard Problem of Responsibility. In D. Shoemaker (Ed.), *Oxford Studies in Agency and Responsibility* (Vol. 3, pp. 160-188). https://doi org.ezproxy.ub.gu.se/10.1093/acprof:oso/9780198744832.003.0009, accessed 8 Nov. 2023.

McKenna, M. (1998). The limits of evil and the role of moral address: A defense of Strawsonian compatibilism. *The Journal of Ethics, 2,* 123-142.

McKenna, M. (2005). Where Frankfurt and Strawson meet. *Midwest Studies in Philosophy, 29*(1), 163-180.

McKenna, M. (2012). *Conversation & Responsibility*. Oxford University Press.

McKenna, M. (2013). Directed Blame and Conversation. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (online edn., pp. 119-140). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199860821.003.0007, accessed 6 Oct. 2023.

McKenna, M. & Coates, D. J. (2021). Compatibilism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/archives/fall2021/entries/compatibilism/

McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life* (online edn). Oxford Academic. https://doi.org/10.1093/0195079981.001.0001, accessed 4 Aug. 2023.

Mehta, A., Kunjadiya, Y., Kulkarni, A., & Nagar, M. (2022). Exploring the viability of Conversational AI for Non-Playable Characters: A comprehensive survey. In D. Kumar, T. Kumar Dey, & S. Dash, *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST*) (pp. 96-102). IEEE.

Meijer, E. (2019). *When animals speak: Toward an interspecies democracy*. New York University Press.

Mele, A. (1995). *Autonomous Agents: From Self-Control to Autonomy.* Oxford University Press.

Mele, A. (1996). Soft Libertarianism and Frankfurt-Style Scenarios. *Philosophical Topics, 24*(2), 123–41.

Mele, A. (2006). *Free Will and Luck*. Oxford University Press.

Mele, A. (2010). Moral Responsibility for Actions: Epistemic and Freedom Conditions. *Philosophical Explorations, 13*(2): 101–111. doi:10.1080/13869790903494556

Mele, A. R. (2019). *Manipulated agents: A window to moral responsibility*. Oxford University Press.

Metzinger, T. (2013). Two principles for robot ethics. In E. Hilgendorf & J.-P. Günther (Eds.), *Robotik und Gesetzgebung* (pp. 247–286). Baden-Baden, Nomos.

Mickelson, K. (2015). The Zygote Argument is invalid: Now what?. *Philosophical Studies, 172,* 2911–2929. https://doi-org.ezproxy.ub.gu.se/10.1007/s11098-015-0449-6

Mickelson, K. (2016). The Manipulation Argument. In M. Griffith, K. Timpe & N. Levy (Eds.), The *Routledge Companion to Free Will* (pp. 166-178). Routledge.

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *Ai & Society, 35*, 957-967.

Mitchell-Yellin, B. (2015). The Platonic model: statement, clarification and defense. *Philosophical Explorations, 18*(3), 378-392.

Monsó, S. (2017). Morality without mindreading. *Mind & Language, 32*(3), 338-357.

Monsó, S. & Andrews, K. (2022). Animal Moral Psychologies. In M. Vargas & J. M. Doris (Eds.), *The Oxford Handbook of Moral Psychology* (Oxford Handbooks, online edn, pp. 388-420). Oxford University Press. https://doi-org.ezproxy.ub.gu.se/10.1093/oxfordhb/9780198871712.013.22, accessed 4 Aug. 2023.

Monsó, S., Benz-Schwarzburg, J., & Bremhorst, A. (2018). Animal morality: What it means and why it matters. *The Journal of Ethics, 22*, 283-310.

Moody-Adams, M. M. (1994). Culture, responsibility, and affected ignorance. *Ethics*, *104*(2), 291-309.

Moore, G. E. (1912). *Ethics*. Clarendon Press

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: a mapping review. *Social Science & Medicine, 260*, 113172.

Mouritsen, H., & Ritz, T. (2005). Magnetoreception and its use in bird navigation. *Current opinion in neurobiology, 15*(4), 406-414.

Müller, V. C. (2016). Autonomous Killer Robots Are Probably Good News. In E. Di Nucci & F. S. de Sio (Eds.), *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons* (pp. 67-81). Routledge.

Müller, V. C. (2021). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford University. https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

Murphy, J. G. (1972). Moral death: A Kantian essay on psychopathy. *Ethics, 82*(4), 284-298.

Murray, S., & Vargas, M. (2020). Vigilance and control. *Philosophical Studies, 177*, 825-843.

Musschenga, A. (2015). Moral animals and moral responsibility. *Les ateliers de l'éthique/The Ethics Forum, 10*(2), 38-59.

Nadler, A., & McGuigan, L. (2018). An impulse to exploit: the behavioral turn in data-driven marketing. *Critical Studies in Media Communication, 35*(2), 151-165.

Nado, J., Kelly, D., Stich, S. (2009). Moral judgment. In J. Symons & P. Calvo (Eds.), *The Routledge Companion to Philosophy of Psychology* (pp. 621-633). Routledge.

Nagel, T. (1974). What is it like to be a bat? *The philosophical review, 83*, 435–450.

Nakajima, S., Arimitsu, K., & Lattal, K. M. (2002). Estimation of animal intelligence by university students in Japan and the United States. *Anthrozoös, 15*(3), 194-205.

Nelkin, D. K. (2011). *Making Sense of Freedom and Responsibility*. Oxford University Press.

Nelkin, D. K. (2015). Psychopaths, Incorrigible Racists, and the Faces of Responsibility, *Ethics, 125*(2): 357–390. doi:10.1086/678372

Nelkin, D. K., & Rickless, S. C. (2017). Moral responsibility for unwitting omissions. In D. Nelkin & S. C. Rickless (Eds.), *The ethics and law of omissions* (pp. 1-32). Oxford University Press.

Newman, J. (2014, Oct 17). To Siri, With Love. *New York Times*.

Nibbeling, N., Oudejans, R. R., Ubink, E. M., & Daanen, H. A. (2014). The effects of anxiety and exercise-induced fatigue on shooting accuracy and cognitive performance in infantry soldiers. *Ergonomics*, *57*(9), 1366-1379.

Nichols, S. (2001). Mindreading and the cognitive architecture underlying altruistic motivation. *Mind & language, 16*(4), 425-455.

Nichols, S. (2002). How psychopaths threaten moral rationalism: Is it irrational to be amoral?. *The Monist, 85*(2), 285-303.

Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgement* (online edn). Oxford University Press. https://doi.org/10.1093/0195169344.001.0001, accessed 7 Aug. 2023.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review, 84*(3), 231.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., ... & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18(*1), 36-88.

Nowell-Smith, P. (1948). Freewill and moral responsibility. *Mind, 57*(225), 45-61.

Nussbaum, M. C. (2005). Beyond "Compassion and Humanity": Justice for Nonhuman Animals. In C. R. Sunstein & M. C. Nussbaum (Eds.), *Animal Rights: Current Debates and New Directions* (online ed., pp. 325-407). Oxford University Press. https://www.jstor.org/stable/j.ctv1c7zftw.10

Nussbaum, M. C. (2006). *Frontiers of Justice: Disability, Nationality, Species Membership*. Harvard University Press. https://doi.org/10.2307/j.ctv1c7zftw

Nyholm, S. (2018). Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics, 24*(4), 1201–1219.

Nyholm, S. (2019). Other minds, other intelligences: The problem of attributing agency to machines. *Cambridge Quarterly of Healthcare Ethics, 28*(4), 592-598.

Nyhom, S. (2020). *Humans and robots: Ethics, agency, and anthropomorphism*. Rowman & Littlefield.

Nyholm, S., & Smids, J. (2020). Can a Robot Be a Good Colleague?. *Science and Engineering Ethics, 26*(4), 2169–2188.

O'Neill, O. (1998). Kant on Duties Regarding Nonrational Nature. *Aristotelian Society Supplementary Volume, 72*(1), 189-210.

O'Connor, T. & Franklin, C. (2022). Free Will. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition). Stanford University. https://plato.stanford.edu/archives/win2022/entries/freewill/

Page, E. A. (2006). *Climate change, justice and future generations*. Edward Elgar Publishing.

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness, 5*(02), 105-129.

Pedersen, H. (2014). Knowledge Production in the 'Animal Turn': Multiplying the Image of Thought, Empathy, and Justice. In E. Andersson Cederholm, A. Björck, K. Jennbert, & A. S. Lönngren (Eds.), *Exploring the animal turn: human-animal relations in science, society and culture* (pp. 13-18). Pufendorf Institute for Advanced Scholars, Lund University.

Peels, R. (2011). Tracing culpable ignorance. *Logos & Episteme, 2*(4), 575-582.

Pelau, C., Dabija, D. C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior, 122*, 106855.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge University Press.

Pereboom, D. (2014a). *Free Will, Agency, and Meaning in Life*. Oxford University Press.

Pereboom, D. (2014b). The disappearing agent objection to event-causal libertarianism. *Philosophical Studies, 169*, 59-69.

Petersen, S., Houston, S., Qin, H., Tague, C., & Studley, J. (2017). The utilization of robotic pets in dementia care. *Journal of Alzheimer's disease, 55*(2), 569-574.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence*. (Report of the 2015-2016 Study Panel). Stanford University. http://ai100.stanford.edu/2016-report, accessed Nov. 8, 2023.

Pickard, H. (2011). Responsibility without blame: Empathy and the effective treatment of personality disorder. *Philosophy, Psychiatry, & Psychology, 18*(3), 209-224.

Pickard, H. (2014). Responsibility without blame: Therapy, philosophy, law. *Prison Service Journal, 213*, 10-16.

Pickard, H. (2015). Psychopathology and the ability to do otherwise. *Philosophy and Phenomenological Research, 90*(1), 135-163.

Pickard, H. (2017). Responsibility without blame for addiction. *Neuroethics, 10*(1), 169-180.

Pickard, H., & Ward, L. (2013). Responsibility without blame: Philosophical reflections on clinical practice. In K. W. M. Fulford et al., (Eds.), *The Oxford Handbook of Philosophy and Psychiatry* (online edn, pp. 1134-1152), https://doi-org.ezproxy.ub.gu.se/10.1093/oxfordhb/9780199579563.013.0066, accessed 15 Nov. 2023.

Powers, T. M. (2013). On the moral agency of computers. *Topoi, 32*(2), 227–236.

Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.

Prinz, J. (2014). Where do morals come from?—a plea for a cultural approach. In M. Christen, M. Huppenbauer, C. Tanner, & C. van Schaik (Eds.), *Empirically Informed Ethics* (pp. 99-116). Springer.

Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.

Putnam, H. (1967). Psychological Predicates. In W. H. Capitan & D.D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37-48). University of Pittsburgh Press.

Rachels, J. (1990). *Created From Animals: The Moral Implications of Darwinism*. Oxford University Press.

Radoilska, L. V. (2023) Moral Competence and Mental Disorder. In M. Kiener (Ed.), *The Routledge Handbook of Philosophy of Responsibility.* Routledge. Retrieved from University of Kent's Academic Repository. https://kar.kent.ac.uk/101448/

Ramsey, W. (2022). Implicit Mental Representation. In J. R. Thompson (Ed.), *The Routledge Handbook of Philosophy and Implicit Cognition* (pp. 33-43). Taylor & Francis.

Regan, T. (2004). *The Case for Animal Rights*. University of California Press. (Original work published 1983)

Reid, T. (1969). *Essays on the Active Powers of the Human Mind*. MIT Press. (Original work published 1788)

Richman, K. A. & Bidshahri, R. (2018). Autism, Theory of Mind and the Reactive Attitudes. *Bioethics 32*, 43–49.

Ritchie, H. (2019) - *Humans make up just 0.01% of Earth's life – what's the rest?*. OurWorldInData.org. https://ourworldindata.org/lifeonearth

Ritchie, H., Spooner, F. & Roser, M. (2022). - *Biodiversity*. OurWorldInData.org. https://ourworldindata.org/biodiversity

Robbins, P. & Jack, A. I. (2006). The phenomenal stance. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *127*(1), 59-85.

Robichaud, P. (2014). On Culpable Ignorance and Akrasia. *Ethics 125*(1), 137–51.

Rosen, G. (2004). Skepticism about Moral Responsibility. *Philosophical Perspectives, 18*, 295–313. doi:10.1111/j.1520-8583.2004.00030.x

Rosen, G. (2015). The alethic conception of moral responsibility. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The nature of moral responsibility: New essays* (pp. 65-88). Oxford University Press.

Rowlands, M. (2012). *Can Animals Be Moral?*. Oxford University Press.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of Applied Learning and Teaching, 6*(1), 342-263.

Rudy-Hiller, F. (2017). A capacitarian account of culpable ignorance. *Pacific Philosophical Quarterly, 98,* 398-426.

Rudy-Hiller, F. (2022a). The Epistemic Condition for Moral Responsibility. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition). Stanford University. https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/

Rudy-Hiller, F. (2022b). Notes to The Epistemic Condition for Moral Responsibility. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition). Stanford University. https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/notes.html

Russell, P. (1992). Strawson's way of naturalizing responsibility. *Ethics, 102*(2), 287-302.

Russell, P. (2002). *Freedom and Moral Sentiment: Hume's Way of Naturalizing Responsibility* (online edn). Oxford University Press. https://doi.org/10.1093/0195152905.001.0001, accessed 9 Jun. 2023.

Russell, P. (2004). Responsibility and the condition of moral sense. *Philosophical Topics, 32*(1/2), 287-305.

Saltik, I., Erdil, D., & Urgen, B. A. (2021). Mind perception and social robots: The role of agent appearance and action types. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 210-214). Association of Computing Machinery.

Sapontzis, S. F. (1980). Are animals moral beings?. *American Philosophical Quarterly, 17*(1), 45-52.

Sapontzis. S. F. (1987). *Morals, Reason, and Animals*. Temple University Press.

Sars, N. (2022). Incapacity, Inconceivability, and Two Types of Objectivity. *Pacific Philosophical Quarterly, 103*(1), 76-94.

Sauppé, A., & Mutlu, B. (2015, April). The social impact of a robot co-worker in industrial settings. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3613-3622). Association for Computing Machinery.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Harvard University Press.

Scanlon, T. M. (2008). *Moral dimensions: Permissibility, meaning, blame*. Harvard University Press.

Scanlon, T. M. (2013). Contractualism and Utilitarianism. In R. Shafer-Landau (Ed.), *Ethical theory: An anthology* (2nd ed, pp. 593-607). John Wiley & Sons. (Reprinted from *Utilitarianism and Beyond*, pp. 103-128, by A. Sen & B. Williams, Eds., 1982, Cambridge University press)

Scarantino, A. & de Sousa, R. (2021). Emotion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition). Stanford University. https://plato.stanford.edu/archives/sum2021/entries/emotion/

Schaerer, E., Kelley, R., & Nicolescu, M. (2009). Robots as animals: A framework for liability and responsibility in human-robot inter- actions. *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 72–77). IEEE.

Schlick, M. (1966). When is a Man Responsible?. In B. Berofsky (Ed.) *Free Will and Determinism* (pp. 54-63). Harper & Row. (Reprinted from *Problems of Ethics*, pp. 143-158, by D. Rynin, Trans., 1939, Prentice-Hall)

Shapiro, K., & DeMello, M. (2010). The state of human-animal studies. *Society & Animals, 18*(3), 307-318.

Shapiro, P. (2006). Moral agency in other animals. *Theoretical Medicine and Bioethics, 27*(4), 357-373.

Shaver, R. (2023). Egoism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition). Stanford University. https://plato.stanford.edu/archives/spr2023/entries/egoism/

Shaw, L. L., Batson, C. D., & Todd, R. M. (1994). Empathy avoidance: Forestalling feeling for another in order to escape the motivational consequences. *Journal of Personality and Social Psychology, 67*(5), 879.

Sher, G. (2009). *Who knew?: Responsibility without awareness.* Oxford University Press.

Sherman, P.W. (1988). The levels of analysis. *Animal Behaviour. 36*, 616–619. doi:10.1016/S0003-3472(88)80039-3

Shettleworth, S. J. (2010). Clever animals and killjoy explanations in comparative psychology. *Trends in cognitive sciences*, *14*(11), 477-481.

Shettleworth, S. J. (2012). Modularity, comparative cognition and human uniqueness. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1603), 2794-2802.

Shoemaker, D. (2003). Caring, identification, and agency. *Ethics, 114*(1), 88-118.

Shoemaker D (2007) Moral address, moral responsibility, and the boundaries of the moral community. *Ethics, 118*(1), 70–108. https://doi.org/10.1086/521280

Shoemaker, D. (2010). Responsibility, agency, and cognitive disability. In E. F. Kittay & L. Carlson (Eds.), *Cognitive disability and its challenge to moral philosophy* (pp. 201-223). Wiley-Blackwell.

Shoemaker, D. (2011). Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility. *Ethics, 121*(3), 602–632. doi:10.1086/659003

Shoemaker, D. (2013). Qualities of will. *Social Philosophy and Policy, 30*(1-2), 95-120.

Shoemaker, D. (2015). *Responsibility from the Margins.* Oxford University Press.

Shoemaker, D. (2017). Response-dependent responsibility; or, a funny thing happened on the way to blame. *Philosophical Review, 126*(4), 481-527.

Shoemaker, D. (2022). Disordered, Disabled, Disregarded, Dismissed: The Moral Costs of Exemptions from Accountability. In M. King & J. May (Eds.), *Agency in Mental Disorder: Philosophical Dimensions* (online ed., pp. 33-62). Oxford University Press. https://doi-org.ezproxy.ub.gu.se/10.1093/oso/9780198868811.003.0003, accessed 9 Aug. 2023.

Sie, M. (2005). *Justifying blame: Why free will matters and why it does not*. Rodopi.

Sie, M. (2014). Self-knowledge and the minimal conditions of responsibility: a traffic-participation view on human (moral) agency. *The Journal of Value Inquiry, 48*(2), 271-291.

Sie, M., & Wouters, A. (2010). The BCN challenge to compatibilist free will and personal responsibility. *Neuroethics, 3*, 121-133.

Simmons, l. & Armstrong, P. (Eds.). (2007). *Knowing Animals*. Leiden & Boston: Brill.

Singer, P. (1975). *Animal liberation: New ethics for our treatment of animals*. Random House

Singer, P. (1993). *Practical ethics* (2nd ed). Cambridge University Press.

Singer, P. (1996). *Rethinking Life and Death: The Collapse of Our Traditional Ethics*. Macmillan.

Sisneros, J. A., & Tricas, T. C. (2002). Neuroethology and life history adaptations of the elasmobranch electric sense. *Journal of Physiology-Paris, 96*(5-6), 379-389.

Skidmore, J. (2001). Duties to animals: The failure of Kant's moral theory. *Journal of Value Inquiry, 35*, 541.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies, 149*, 102601.

Sliwa, P. (2016). Moral Worth and Moral Knowledge. *Philosophy and Phenomenological Research, 93*, 393-418. https://doi.org/10.1111/phpr.12195

Smart, J. J. C. (1961). Free-will, praise and blame. *Mind, 70*(279), 291-306.

Smith, A. M. (2005). Responsibility for Attitudes: Activity and Passivity in Mental Life, *Ethics, 115*(2): 236–271. doi:10.1086/426957

Smith, A. M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies, 138*, 367-392.

Smith, A. M. (2012). Attributability, answerability, and accountability: In defense of a unified account. *Ethics*, *122*(3), 575-589.

Smith, H. M. (1983). Culpable Ignorance, *Philosophical Review, 92*(4), 543–71. doi:10.2307/2184880

Sneddon, A. (2005). Moral responsibility: The difference of Strawson, and the difference it should make. *Ethical theory and moral practice, 8*, 239-264.

Sommers, T. (2007). The objective attitude. *The Philosophical Quarterly, 57*(228), 321-341.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies, 173*, 1203-1232.

Sripada, C, & Stich, S.P. (2007). A Framework for the Psychology of Norms. In P. Carruthers , S. Laurence, & S. Stich (Eds.), *The Innate Mind Volume 2: Culture and Cognition* (pp. 280-301). Oxford University Press.

Stenning, A. (2020). Understanding empathy through a study of autistic life writing. On the importance of neurodivergent morality. In H. Bertilsdotter Rosqvist, N. Chown, & A. Stenning (Eds.), *Neurodiversity studies. A new critical paradigm* (pp. 108–124). Routledge.

Stich S (2009) Replies. In D. Murphy & M. Bishop (Eds.), *Stich and his critics* (pp. 190-252). Wiley.

Stieglitz, S., Brachten, F., Ross, B., & Jung, A. K. (2017). Do social bots dream of electric sheep? A categorisation of social media bot accounts. *Proceedings of the Australasian Conference on Information Systems*, *89*.

Stout, N. (2016). Conversation, responsibility, and autism spectrum disorder. *Philosophical Psychology, 29*(7), 1015-1028. 10.1080/09515089.2016.1207240

Strawson, G. (2010) *Freedom and Belief* (online ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199247493.001.0001, accessed 3 Aug. 2023.

Strawson, P. F. (1982). Freedom and Resentment. In G. Watson (Ed.), *Free Will* (pp. 59-80). Oxford University Press. (Reprinted from "Freedom and Resentment", 1962, *Proceedings of the British Academy, 48*, 187-211)

Strawson, P. F. (1985). *Skepticism and Naturalism: Some Varieties*. Columbia University Press.

Stump, E. (1988). Sanctification, Hardening of the Heart, and Frankfurt's Concept of Free Will. *Journal of Philosophy, 85*: 395–420.

Sullins, J. P. (2011). When is a robot a moral agent. *Machine ethics, 6*(2001), 151-161.

Sullivan, Y. W., & Fosso Wamba, S. (2022). Moral judgments in the age of artificial intelligence. *Journal of Business Ethics, 178*(4), 917-943.

Sunstein, C. R. (2005). Can Animals Sue?. In C. R. Sunstein & M. C. Nussbaum (Eds.), *Animal Rights: Current Debates and New Directions* (online ed., pp. 251-262). Oxford University Press. https://doi-org.ezproxy.ub.gu.se/10.1093/acprof:oso/9780195305104.003.0012, accessed 22 Nov. 2023.

Sytsma, J., & Machery, E. (2012). The two sources of moral standing. *Review of Philosophy and Psychology*, *3*, 303-324.

Szigeti, A. (2012). Revisiting Strawsonian arguments from inescapability. *Philosophica, 85*(2), 91-121.

Talbert, M. (2006). Contractualism and our duties to nonhuman animals. *Environmental Ethics, 28*(2), 201-215.

Talbert, M. (2008). Blame and responsiveness to moral reasons: Are psychopaths blameworthy?. *Pacific Philosophical Quarterly, 89*(4), 516-535.

Talbert, M. (2012). Moral competence, moral blame, and protest. *The Journal of Ethics, 16*, 89-109.

Talbert, M. (2013). Unwitting wrongdoers and the role of moral disagreement in blame. In D. Shoemaker & N. A. Tognazzini (Eds.), *Oxford Studies in Agency and Responsibility* (Vol. 1, pp. 225-245). Oxford University Press.

Talbert, M. (2017). Akrasia, awareness, and blameworthiness. In P. Robichaud & J. W. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 47-63). Oxford University Press.

Talbert, M. (2021). Psychopaths and Symmetry: A Reply to Nelkin. *Philosophia 49* (3), 1233-1245.

Talbert, M. (2022). Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Ed.). Stanford University. https://plato.stanford.edu/archives/fall2022/entries/moral-responsibility/

Taylor, R. (1963). *Metaphysics*. Englewood Cliffs, Prentice Hall.

Taylor, K., & Alvarez, L. R. (2019). An estimate of the number of animals used for scientific purposes worldwide in 2015. *Alternatives to Laboratory Animals, 47*(5-6), 196-213.

Terada, K., Shamoto, T., Ito, A., & Mei, H. (2007). Reactive movements of non-humanoid robots cause intention attribution in humans. *2007 IEEE/RSJ international conference on intelligent robots and systems* (pp. 3715-3720). IEEE.

Thellman, S., de Graaf, M., & Ziemke, T. (2022). Mental state attribution to robots: A systematic review of conceptions, methods, and findings. *ACM Transactions on Human-Robot Interaction (THRI), 11*(4), 1-51.

Thellman, S., Giagtzidou, A., Silvervarg, A., & Ziemke, T. (2020). An implicit, non-verbal measure of belief attribution to robots. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 473-475). ACM.

Tiboris, M. (2014). Blaming the Kids: Children's Agency and Diminished Responsibility. *Journal of Applied Philosophy, 31*(1), 77-90.

Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology, 34*(3), 589-607.

Timpe, K. (2011). Tracing and the Epistemic Condition on Moral Responsibility. *Modern Schoolman, 88*(1–2): 5–28. 10.5840/schoolman2011881/22

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift für Tierpsychologie, 20*, 410-433. https://doi.org/10.1111/j.1439-0310.1963.tb01161.x

Tognazzini, N. A. (2015). The strains of involvement. In R. Clarke, M. McKenna, & A. M. Smith (Eds.), *The Nature of Moral Responsibility* (pp. 19-44). Oxford University Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly review of biology*, *46*(1), 35-57.

Tulving, E. (2005). Episodic Memory and Autonoesis: Uniquely Human? In H.S. Terrace & J. Metcalfe (Eds.), *The Missing Link in Cognition: Origins of Self-Reflective Consciousness* (pp. 3-56). Oxford University Press.

Tye, M. (2021). Qualia. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 Ed.). Stanford University. https://plato.stanford.edu/archives/fall2021/entries/qualia/

Uhlmann, E. L., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16*(6), 474-480.

Umbrello, S., Torres, P., & De Bellis, A. F. (2020). The future of war: could lethal autonomous weapons make conflict more ethical?. *AI & SOCIETY, 35*, 273-282.

Van Gulick, R. (2022). Consciousness. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Ed.). Stanford University. https://plato.stanford.edu/archives/win2022/entries/consciousness/

van Inwagen, P. (1975). The Incompatibility of Free Will and Determinism. *Philosophical Studies, 27*, 185–99.

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford University Press.

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*, 719-735.

Vargas, M. (2013). *Building Better Beings: A theory of moral responsibility*. Oxford University Press.

Vargas, M. (2018). The Social Constitution of Agency and Responsibility: Oppression, Politics, and Moral Ecology. In K. Hutchison, C. Mackenzie, & M. Oshana (Eds.), *Social Dimensions of Moral Responsibility* (pp. 110-136). Oxford University Press. https://doi-org.ezproxy.ub.gu.se/10.1093/oso/9780190609610.003.0005, accessed 8 Nov. 2023.

Vargas, M. (2022). Instrumentalist Theories of Moral Responsibility. In D. K. Nelkin & D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (online ed., pp. 3-26). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190679309.013.13, accessed 3 Aug. 2023.

Varner, G. (2001). Sentientism. In D. Jamieson (Ed.), *A Companion to Environmental Philosophy* (pp. 192-203). Blackwell.

Varner, G. (2012). *Personhood, ethics, and animal cognition: Situating animals in Hare's two level utilitarianism*. Oxford University Press.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Veruggio, G. (2006). The euron roboethics roadmap. In *2006 6th IEEE-RAS international conference on humanoid robots* (pp. 612-617). IEEE.

Vihvelin, K. (2022). Arguments for Incompatibilism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Ed.). Stanford University. https://plato.stanford.edu/archives/fall2022/entries/incompatibilism-arguments/

Vincent, S., Ring, R., & Andrews, K. (2018). Normative practices of other animals. In A. Zimmerman, K. Jones, & M. Timmons (Eds.), *The Routledge handbook of moral epistemology* (pp. 57-83). Routledge.

von Rohr, R. C., Burkart, J. M., & Van Schaik, C. P. (2011). Evolutionary precursors of social norms in chimpanzees: a new approach. *Biology & Philosophy, 26*, 1-30.

von Uexküll, J. (2010). *A Foray into the Worlds of Animals and Humans with a Theory of Meaning.* University of Minnesota Press. (Original work published 1934)

Wallace, R. J. (1994). *Responsibility and the moral sentiments.* Harvard University Press.

Waller, B. N. (1997). What rationality adds to animal morality. *Biology and Philosophy, 12*, 341-356.

Waller, B. N. (2006). Denying responsibility without making excuses. *American Philosophical Quarterly, 43*(1), 81-90.

Warren, M A. (1995). Postscript on Infanticide. In G. Percesepe (Ed.), *Introduction to Ethics: Personal and Social Responsibility in a Diverse World* (p. 453). Prentice-Hall.

Warren, M A. (1997). *Moral status: Obligations to persons and other living things*. Clarendon Press.

Wasserman, D, Asch, A., Blustein, J., & Putnam, D. (2017). Cognitive Disability and Moral Status. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017 Ed.). Stanford University. https://plato.stanford.edu/archives/fall2017/entries/cognitive-disability/

Watson, G. (1975). Free agency. *The Journal of Philosophy, 72*(8), 205-220.

Watson, G. (2004). Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. In G. Watson (Ed.), *Agency and Answerability: Selected Essays* (online ed, pp. 219–259). Oxford University Press. doi:10.1017/CBO9780511625411.011, accessed 3 Aug. 2023. (Reprinted from *Responsibility, Character and the Emotions*, pp. 256-286, by F. D. Schoeman, Ed., 1987, Cambridge University Press)

Watson, G. (2004). Two Faces of Responsibility. In G. Watson (Ed.), Agency and Answerability: Selected Essays (online ed, pp. 219-259). Oxford University Press. https://doi-org.ezproxy.ub.gu.se/10.1093/acprof:oso/9780199272273.003.0010, accessed 8 Nov. 2023. (Reprinted from "Two Faces of Responsibility", 1996, *Philosophical Topics, 24*(2), 227–248.

Watson, G. (2011). The Trouble with Psychopaths. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon* (online ed, pp. 307-332). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199753673.003.0013, accessed 5 Aug. 2023.

Watson, G. (2013). XIV—Psychopathic agency and prudential deficits. *Proceedings of the Aristotelian Society, 113*(3, 3), 269-292.

Watson, G. (2014). Peter Strawson on responsibility and sociality. In D. Shoemaker & N. Tognazzini (Eds.), *Oxford Studies in Agency and Responsibility* (vol. 2, pp. 15-32). Oxford University Press.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113-117.

Weatherson, B. (2019). *Normative externalism*. Oxford University Press.

Westra, E., & Andrews, K. (2022). A pluralistic framework for the psychology of norms. *Biology & Philosophy, 37*(5), 40.

Wiggins, D. (1973). Towards a Reasonable Libertarianism. In T. Honderich & P. Kegan (Eds.), *Essays on Freedom of Action* (pp. 31-62). Routledge.

Wilson, T. (2002). *Strangers to ourselves: Discovering the adaptive unconscious.* Harvard University Press.

Wolf, S. (1981). The importance of free will. *Mind, 90*(359), 386-405.

Wolf, S. (1990). *Freedom Within Reason*. Oxford University Press.

Wolf, S. (2011). Blame, Italian Style. In R. J. Wallace, R. Kumar, & S. Freeman (Eds.), *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon* (online ed, pp. 332-347). Oxford University Press. https://doi-org.ezproxy.ub.gu.se/10.1093/acprof:oso/9780199753673.003.0014, accessed 9 Aug. 2023.

Wolf, S. (2013). Sanity and the Metaphysics of Responsibility. In R. Shafer-Landau, R. (Ed.), *Ethical theory: An anthology* (2nd ed., pp. 330-339). John Wiley & Sons. (Reprinted from *Responsibility, Character, and the Emotions*, pp. 46-62, by F. D. Schoeman, Ed., 1987, Cambridge University Press)

Wolf, S. (2015). Responsibility, moral and otherwise. *Inquiry, 58*(2), 127-142.

Wood, A. W. (1998). Kant on duties regarding nonrational nature. *Aristotelian Society Supplementary Volume 72*(1), 189-210.

Wood, A. W. (1999). *Kant's Ethical Thought*. Cambridge University Press.

Wrage, B. (2022). Caring animals and care ethics. *Biology & Philosophy, 37*(3), 18.

Xu, Y., Shieh, C. H., van Esch, P., & Ling, I. L. (2020). AI customer service: Task complexity, problem-solving ability, and usage intention. *Australasian Marketing Journal, 28*(4), 189-199.

Zaki, J. (2018). Empathy is a moral force. In K. Gray, & J. Graham (Eds.), *Atlas of Moral Psychology* (pp. 49-58). The Guilford Press.

Zimmerman, M. J. (1997). Moral responsibility and ignorance. *Ethics, 107*(3), 410-426.

Zimmerman, M. J. (2002). Controlling ignorance: A bitter truth. *Journal of Social Philosophy, 33*(3), 483-490.

Svensk sammanfattning

Kan icke-mänskliga djur och artificiell intelligens (AI) tillskrivas moraliskt agens? Inom filosofin har den rådande uppfattningen varit att endast människor är moraliska agenter då de ensamma besitter fri vilja och förmåga till medveten reflektion. Denna uppfattning har resulterat i att endast människor anses vara moraliskt ansvariga för sina handlingar och därmed som föremål för moraliskt beröm och klander. Trots att djur och maskiner kan orsaka skada, har de därför hittills inte ansetts vara moraliskt ansvariga.

Denna avhandling ifrågasätter den skarpa gränsdragningen mellan människor och icke-mänskliga varelser. Istället för det traditionella synsättet där moraliskt agens anses kräva särskilda inre individuella förmågor tillämpas ett alternativt förhållningssätt där moraliskt agens förstås som förmåga till deltagande i sociala ansvarspraktiker. Genom att skifta fokus från inre individuella egenskaper till en kontextuell och socialt avhängig färdighet verkar det *praktik-fokuserade* angreppssättet göra det mer rimligt att tala om moraliskt agens hos ickemänskliga djur och AI.

Utifrån den nuvarande såväl som troliga framtida förekomsten av både ickemänskliga djur och AI i mänskliga sociala sammanhang så skulle en utvidgning av vem som är moralisk agent potentiellt kunna leda till en radikal förändring av våra sociala, moraliska och juridiska praktiker. Denna avhandling hävdar att de möjliga konsekvenserna av en sådan utvidgning är viktiga att beakta och föreslår att frågan om huruvida moraliskt agens kan tillskrivas djur eller maskiner därför borde angripas som en normativ, *bör*-fråga, istället för som en rent teoretisk, *är*-fråga.

ACTA PHILOSOPHICA GOTHOBURGENSIA
ISSN 0283-2380

Subscriptions to the series and orders for individual copies are sent to:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222, 405 30 Göteborg, Sweden
acta@ub.gu.se

Volumes published:

1. MATS FURBERG, THOMAS WETTERSTRÖM & CLAES ÅBERG (eds.). *Logic and abstraction: Essays dedicated to Per Lindström on his fiftieth birthday*. Göteborg 1986

2. STAFFAN CARLSHAMRE. *Language and time: An attempt to arrest the thought of Jacques Derrida.* Göteborg 1986

3. CLAES ÅBERG (ed.). *Cum grano salis: Essays dedicated to Dick A. R. Haglund*. Göteborg 1989

4. ANDERS TOLLAND. *Epistemological relativism and relativistic epistemology*. Göteborg 1991

5. CLAES STRANNEGÅRD. *Arithmetical realizations of modal formulas*. Göteborg 1997

6. BENGT BRÜLDE. *The Human Good.* Göteborg 1998

7. EVA MARK. *Självbilder och jagkonstitution*. Göteborg 1998

8. MAY TORSETH. *Legitimate and Illegitimate Paternalism in Polyethnic Conflicts*. Göteborg 1999

9. CHRISTIAN MUNTHE. *Pure Selection: The Ethics of Preimplantation Genetic Diagnosis and Choosing Children without Abortion*. Göteborg 1999

10. JOHAN MÅRTENSSON. *Subjunctive Conditionals and Time: A Defense of a Weak Classical Approach*. Göteborg 1999

11. CLAUDIO M. TAMBURRINI. *The "Hand of God"? Essays in the Philosophy of Sports*. Göteborg 2000

12. LARS SANDMAN. *A Good Death: On the Value of Death and Dying*. Göteborg 2001

13. KENT GUSTAVSSON. *Emergent Consciousness: Themes in C.D. Broad's Philosophy of Mind.* Göteborg 2002

14. FRANK LORENTZON *Fri Vilja?* Göteborg 2002

15. JAN LIF. *Can a Consequentialist be a real friend? (Who cares?)* Göteborg 2003

16. FREDRIK SUNDQVIST. *Perceptual Dynamics: Theoretical foundations and philosophical implications of gestalt psychology*. Göteborg 2003

17. JONAS GREN. *Applying utilitarianism: The problem of practical action-guidance*. Göteborg 2004

18. NIKLAS JUTH. *Genetic Information Values and Rights: The Morality of Presymptomatic Genetic Testing*. Göteborg 2005

19. SUSANNA RADOVIC. *Introspecting Representations*. Göteborg 2005

20. PETRA ANDERSSON. *Humanity and nature: Towards a consistent holistic environmental ethics*. Göteborg 2007

21. JAN ALMÄNG. *Intentionality and intersubjectivitiy*. Göteborg 2007

22. ALEXANDER ALMÉR. *Naturalising intentionality: Inquiries into realism & relativism*. Göteborg 2007

23. KRISTOFFER AHLSTRÖM. *Constructive Analysis: A Study in Epistemological Methodology*. Göteborg 2007

24. RAGNAR FRANCÉN. *Metaethical Relativism: Against the Single Analysis Assumption*. Göteborg 2007

25. JOAKIM SANDBERG. *The Ethics of Investing: Making Money or Making a Difference*? Göteborg 2008

26. CHRISTER SVENNERLIND. *Moderate Nominalism and Moderate Realism*. Göteborg 2008

27. JÖRGEN SJÖGREN. *Concept Formation in Mathematics*. Göteborg 2011

28. PETER GEORGSSON. *Metaphor and Indirect Communication in Nietzsche*. Göteborg 2014

29. MARTIN KASÅ. *Truth and Proof in the Long Run: Essays on Trial-and-Error Logics*. Göteborg 2017

30. RASMUS BLANCK. *Contributions to the Meta-mathematics of Arithmetic: Fixed Points, Independence, and Flexibility*. Göteborg 2017

31. MARTIN FILIN KARLSSON. *All There Is: On the semantics of Quantification over Absolutely Everything*. Göteborg 2018

32. PAUL KINDVALL GORBOW. *Self-Similarity in the Foundations*. Göteborg 2018

33. YLWA SJÖLIN WIRLING. *Modal Empiricism Made Difficult: An Essay in the Meta-Epistemology of Modality*. Göteborg 2019

34. MARCO TIOZZO. *Moral Disagreement and the Significance of Higher-Order Evidence*. Göteborg 2019

35. ALVA STRÅGE. *Minds, Brains and Desert. On the Relevance of Neuroscience for Retributive Punishment*. Göteborg 2019

36. STELLAN PETERSSON. *Disarming Context Dependence. A Formal Inquiry into Indexicalism and Truth-Conditional Pragmatics*. Göteborg 2019

37. THOMAS HARTVIGSSON. *Explorations of the Relationship Between the Right to Make Decisions and Moral Responsibility in Healthcare*. Göteborg 2019

38. ALEXANDER ANDERSSON *Giving Executives Their Due: Just Pay, Desert, and Equality*. Göteborg 2021

39. LEILA EL-ALTI *Confluence and Divergence of Emancipatory Healthcare Ideals and Psychiatric Contextual Challenges*. Göteborg 2022

40. RICHARD ENDÖRFER *Weapons of Mass Destruction: Financial Crises from a Philosophical Perspective*. Göteborg 2022

41. DORNA BEHDADI *Nonhuman Moral Agency: A Practice-Focused Exploration of Moral Agency in Nonhuman Animals and Artificial Intelligence*. Göteborg 2023

ACTA PHILOSOPHICA GOTHOBURGENSIA 41

Can nonhuman animals and artificial intelligence (AI) entities be attributed moral agency? The general assumption in the philosophical literature is that moral agency applies exclusively to humans since they alone possess free will or capacities required for deliberate reflection. Consequently, only humans have been taken to be eligible for ascriptions of moral responsibility in terms of, for instance, blame or praise, moral criticism, or attributions of vice and virtue. Animals and machines may cause harm, but they cannot be appropriately ascribed moral responsibility for their behavior.

This thesis challenges the conventional paradigm by proposing an alternative approach where moral agency is conceived as the competence to participate in moral responsibility practices. By shifting focus from intra-individual to contextual and socially situated features, this *practice-focused* approach appears to make the attribution of moral agency to nonhuman animals and AI entities more plausible than commonly assumed.

Moreover, considering the current and potential future prevalence of nonhuman animals and AI entities in everyday settings and social contexts, a potential extension of moral agency to such entities could very well transform our social, moral, and legal practices. Hence, this thesis proposes that the attribution or withholding of moral agency to different entities should be carefully evaluated, considering the potential normative implications.

Dorna Behdadi holds degrees in biology (ethology) and philosophy. Their research interests include moral psychology, moral agency and responsibility, philosophy of mind, comparative cognition, machine ethics and animal ethics.

ISBN 978-91-7963-149-9
9789179631499

ACTA UNIVERSITATIS GOTHOBURGENSIS