Introduction

Oisín Deery[†]
oisin@oisindeery.com

I am delighted to introduce this special issue on the topic of "New work on agency and responsibility." Its aim is to showcase a variety of fresh, new approaches to the topics of agency, free will, and moral responsibility. The issue is divided into five sections, respectively on methodology, the phenomenology of agency, legal and moral responsibility, science and agency, and artificial agency. The titles of these sections derive from my own interests in agency and my inclination to think that philosophical theorizing about free agency and moral responsibility may be due an overhaul. Methodologically, for example, I have argued in my own work that philosophers of free will should take naturalistic demands on their theorizing more seriously. In short, our best theories of free will and responsibility shouldn't be beholden to what we intuitively think the conditions are for these phenomena. Instead, theories of agency and responsibility must be beholden to empirical findings and suitable a posteriori theoretical commitments, such that the phenomena themselves are things we can make *discoveries* about. As such, we might find that being free or responsible is different from what we thought it was. The papers in Section 1, whose authors have deeply informed my own thinking, make similar claims.

I have also argued that our best theory of what it is to act freely ought to do justice to our free-agency phenomenology, but only in a deflationary way that provides support for a naturalistic theory. In Section 2, Terence Horgan and Mark Timmons argue for a strong version of this claim.

If anything like the naturalistic approach to agency and responsibility that I (and a growing number of others) believe is right, then we, as philosophical theorizers, must be clearer about how empirical work should inform our theorizing, and about the normative demands on our theories. In the latter regard, it's no longer tenable that a simplistic view, unmoored from the untidy workings of our actual responsibility practices, will suffice. In their own ways, the papers

[†] Macquarie University, Australia & York University, Canada.

in Section 1 already address this issue, and the papers in Section 3, on "Moral and legal responsibility," pick up on this theme as well. They do so by looking more carefully at the often messy, interlocking varieties of control that underpin our apparently free and responsible actions, at the practicalities of how our judgments of legal responsibility are informed by (and often detached from) our judgments of moral responsibility, and at whether punishment is as easily justifiable as we think, even if we are in fact agents of the sort we think we are. It's heady stuff.

Likewise, for someone like myself who thinks that free agency, and our concept of free agency, is amenable to a thoroughly naturalistic approach, it's of paramount importance to ask how and why we acquired the abilities that underpin our free agency, as well as how and why we acquired its related concept – and other adjacent concepts – in our developmental learning and also evolutionarily. The papers in Section 4, on "Science and agency," point us in the sorts of directions we should, in my view, be headed. These papers examine human infants' developing notions of agency and free will and the existence of joint intentions in great apes' social cognition.

Section 5, on "Artificial moral agency," grapples with the question of how we ought to think about the artificial moral agents we might one day create. This has been an emerging interest in my own work, including in my collaborative work with Katherine Bailey (e.g., Deery and Bailey, 2022).

So, with apologies to the authors, I'll come clean. I invited all of these authors to contribute papers (which were, of course, thoroughly and anonymously peer-reviewed) in the hope that their contributions would inform my own thinking about these issues. And wow, did they ever deliver!

With this admission out of the way, I'll start getting myself out of the way too. In the rest of this introduction, I will briefly introduce the papers before getting entirely out of the way and letting these wonderful contributions speak to you on their own, and in their own terms.

1. Methodology and free will

In her paper, "A discretionary case for preservationism about free will," Kelly McCormick addresses the question of what we should do if empirical findings, for example, undermine our intuitive notion of free agency. Should we preserve this notion or eliminate it? McCormick defends a particular variety of preservationism

about the notion of free agency. In this paper, she does so, first, by defending Shaun Nichols' preservationist view against a criticism of Gregg Caruso's.

Nichols (2015, 2017, forthcoming) maintains that our answer to the question of whether we actually have free will depends on the reference convention that we adopt for the relevant term or concept. On a descriptivist convention, roughly, the term 'free will' successfully refers only if (most of) the intuitive presuppositions that we associate with the term are satisfied by some actual humans. The worry for Nichols – and others – is that it seems increasingly unlikely that (enough of) these presuppositions are actually satisfied. If so, eliminativism about free will would win. In other words, we should eliminate the term and conclude that free will does not exist.

But we don't *have* to do that. Why? Because reference for the term or concept might be fixed *not* by a descriptivist convention but instead by (for example) a causal-historical convention. Accordingly, reference would succeed as long as the paradigm cases of free will are held together in the right way. In that case (where much of the work involves unpacking "the right way"), we might discover that free will is different than we thought. That is, free will might exist even if our intuitive presuppositions about it aren't satisfied by any human beings. As long as the requirements of the causal-historical convention are met, preservationism about free will succeeds.

Nichols himself advocates *pluralism* about reference, maintaining that it's systematically ambiguous whether 'free will' refers: under a plausible non-descriptivist and preservationist reference convention, it probably does refer; yet under an equally plausible descriptivist and eliminativist convention, it might not. Additionally, Nichols recommends *discretionism* about free will, according to which we should strategically adopt the preservationist convention in some contexts, yet the eliminativist one in others, depending on the particular context and our practical interests at that time (2015: 74). However, Caruso (e.g., 2015) maintains that on Nichols' own theoretical commitments, this pluralist view collapses into eliminativism, and thus both pluralism and discretionism are false (since the latter position depends on the former one).

In my own recent work, I've argued that Caruso's criticism of Nichols fails (Deery, 2021b: 137–156). McCormick takes a different, though complementary, tack in defending the preservationist part of Nichols' view against Caruso. But that's not where McCormick's paper truly shines. In the second part of her paper, McCormick instead sets her sights on Nichols' own view.

McCormick maintains that what she calls the "motivating moral concerns" (2022: 16–25) behind both eliminativism and preservationism are distinct and important to be clear about and acknowledge. The motivating concern behind eliminativist views like Caruso's, McCormick thinks, is a concern about undeserved harm. After all, blaming and punishment harm, and if no one has free will then no one deserves to be harmed by blame or punishment. By contrast, McCormick thinks that the concern behind preservationist views (e.g., Heller, 1996; Vargas, 2013, forthcoming; Nichols, 2015; Deery, 2021a, 2021b) is to ensure that we *can* and *do* attribute free will, as well as moral and legal responsibility, in cases where agents wrong one another (intentionally, etc.). In those cases, there are victims of a wrong, and to fail to hold the wrongdoer morally responsible (as the eliminativist would counsel) is to fail, McCormick claims, to protect and defend the interests of these victims. So, the reference convention that we adopt for 'free will' can't just be entirely up to our "discretion." *Pace* Nichols' claims, we must decide.

> When adjudicating between preservationists and eliminativists, we ought to pay closer attention to the moral concerns that *motivate* each of these positions, more clearly *articulate* the role that these concerns play in determining the reference of 'free will', *acknowledge* the moral significance of both kinds of concerns, and ultimately take a stand on which moral concern *we should care about more*. And I submit that we should, all-things-considered, care more about protecting and defending victims than avoiding the potentially undeserved harms for wrongdoers embedded in our responsibility-related attitudes and practices. (McCormick 2022: 24)

McCormick thinks that while a pluralist, discretionist variety of preservationism is secure against eliminativist criticisms like Caruso's, outright preservationism should be preferred to discretionism.

In the second paper of Section 1, Henry Argetsinger and Manuel Vargas focus on a related problem about responsibility. Their paper, "What's the relationship between the theory and practice of moral responsibility?" argues that while many attributions of responsibility are likely successful,

> ... they are subject to a range of failures as well. In particular, from the standpoint of going theories of responsibility, there is a wide range of cases where our actual attributions of responsibility rely on properties and concepts not identified by philosophical theories of responsibility. (2022: 37)

The problem, as Argetsinger and Vargas see it, is as follows. The fact that our judgments or attributions of responsibility are "polluted" in this way by "heuristics and biases" might be taken by some theorists of responsibility as encouragement to revise away from ordinary talk and thought *in order* to achieve extensional accuracy. But such a theory will, to the extent that it's revisionary in this way, lose normative authority. Conversely, however, to the extent that a theory cleaves to ordinary talk and thought at the *expense* of extensional accuracy, it's unsatisfying too. Argetsinger and Vargas claim that any satisfactory theory of responsibility must fist aim at and achieve extensionally accuracy, in being grounded in properties actually found in the relevant cases. Yet it must *also* have what they call "normative authority," in that it should be "grounded in a satisfying theory of the normative basis of responsibility practices" (2022: 32).

As a result, our theory and practice pull in opposite directions. Argetsinger and Vargas conclude that...

> ... our ambition has been to call greater attention to an underappreciated methodological challenge for most existing theories of responsibility, one that becomes particularly visible when we consider the role of heuristics and biases in responsibility attributions. The most obvious ways of responding to the divergence between theory and practice each raise non-trivial challenges for a metaphysical theory of moral responsibility that attempts to do without some account of the normative grounds of responsibility. If we are right, a satisfying theory of responsibility will give us both an account of the metaphysical and normative foundations of responsibility, and this is the only way to address the tension generated by pressures for extensional accuracy and normative guidance. In sum, metaphysics is not enough. (2022: 56)

In classic philosophical fashion, Argetsinger and Vargas leave us with a problem rather than any clear answer. But no theory of responsibility can be satisfying, they say, unless it addresses this problem.

2. Phenomenology

Related to the questions about the referential success or extensional accuracy of our terms and theories is the question of how we acquire the concept of free will. Focusing on this question, it seems plausible that we acquire our concept not simply as a result of our tracking certain properties in other agents (perhaps often for the purpose of attributing responsibility) but also as a result of our own possession of (at least some subset of) these properties. Regarding this possibility, a prominent suggestion, historically, has been that we acquire the concept

of free will partly as a result of how we experience our own apparently free agency (e.g., Strawson, 1986; cf. Deery et al., 2013; Nichols, 2015; Deery, 2015; Horgan, 2015). However, even this seemingly straightforward suggestion is fraught with difficulty and raises a number of further questions.

For instance, it's not always clear what the relevant phenomenology is, or whether we could reliably introspect it in such a way that it would be relevant to answering questions about the referential success of the term or concept of free will, or the extensional accuracy of our theories of responsibility. Worse, even *reliable* introspection might not justify certain answers to these questions. Finally, even if consideration of phenomenology did help in addressing *some* questions about referential success, it might not address questions about the extensional accuracy of our theories of responsibility – i.e., the question most philosophers are interested in.

Even so, phenomenology may, as some have suggested, be a more basic source of many people's understanding of free will and their concern about whether we have it. Moreover, these difficulties have not stopped philosophers from maintaining that our experiences as of acting freely might help to fix the reference of our term or concept (e.g., Caruso, 2015). This question has motivated some philosophers who are skeptical about the existence of free will, since they take scientific discoveries to reveal that we lack the sort of agency that we experience possessing. Yet it has also motivated philosophers who maintain that we *have* free will. For example, Terence Horgan has long maintained (e.g., Horgan, 2007; 2011; 2014; 2015) that it's plausible that the reference conditions of our term or concept of free will might be largely inherited from the satisfaction conditions of our free-agency phenomenology. If the phenomenology is libertarian, as some think, then it's not only inaccurate if determinism is true, and is thus libertarian, but the concept or term may also inherit libertarian satisfaction conditions from the relevant phenomenology. Free will wouldn't exist unless libertarianism is true.

Horgan defends compatibilism in response to both of these worries. He has done so by developing an error theory for libertarian judgments about free-agency experiences. Accordingly, even if people tend to *judge* their phenomenology as being libertarian in having incompatibilist satisfaction conditions, actually it is compatibilist: people *misinterpret* their phenomenology. So, even assuming determinism, the phenomenology might be accurate, and if the concept inherits its reference conditions from the phenomenology, the concept can still refer.

Horgan and Mark Timmons pick up on this debate in their paper in Section 2, "Is agentive freedom a secondary quality?" In it, they present an argument against the claim that agentive phenomenology has libertarian content (cf. Deery et al., 2013). Horgan and Timmons endorse a thoroughgoing compatibilist account of the satisfaction conditions of free-agency phenomenology that also applies to judgments about free agency – a view they call *uniform compatibilism*. In short, they argue that phenomenology does *not* have introspectible libertarian content, in the way that most libertarians and even some compatibilists hold. For example, according to what Horgan and Timmons call *illusionist compatibilism*, the illusory component of free-agency phenomenology is a secondary quality that has two distinct varieties of veridicality or accuracy conditions, one of which couldn't be satisfied if determinism is true yet the other of which *might* be satisfied (as Horgan and Timmons correctly note, this view has been developed and presented, though not endorsed, by me (e.g., Deery, 2015; 2021b)). Horgan and Timmons maintain that their view has a number of advantages over illusionist compatibilism. As a result, and in answer to the question that they pose in the title of their paper, Horgan and Timmons conclude that, "No, agentive freedom is not a secondary quality" (2022: 86).

3. Moral and legal responsibility

David Shoemaker opens Section 3 in a characteristically provocative way in his paper, "Empathic control." In an epigraph, Shoemaker quotes Michael McKenna, who writes that "[A] unifying requirement on moral responsibility is that control comes in *somewhere*" (2008: 36). But *what* sort of control? On the standard view, according to Shoemaker, moral responsibility requires *voluntary* control. And while the "nature of the voluntary is somewhat obscure, ... all ... agree on the object of voluntary governance, namely, actions, either physical or mental" (2022: 91).

Other philosophers have argued, however, that agents also seem responsible for things that *aren't* actions and *aren't* under voluntary control. For example, they claim that we are responsible for at least some of our *attitudes*, but not because we exercise voluntary control over them (because we can't). According to these theorists, we legitimately blame people, sometimes, for their attitudes. If I judge that a friend is worth caring about, but I forget their birthday, the attitude reflected in my actual behavior reveals that "I don't care as much as I claim I do" (Shoemaker, 2022: 93), and I am criticizable for this behavior and the attitude that it reveals.

As a result,

> Control must come in two flavors: Volitional control... governs the *actions* for which agents are responsible, and evaluative control governs the *attitudes* for which agents are responsible. (Shoemaker, 2022: 95)

But if we open the door to evaluative control, Shoemaker thinks we also have reason to admit another form of control, since there is "an additional psychic stance for which we are often held responsible, even though it is not governed by either volitional or evaluative control" (2022: 95). For Shoemaker, this further stance is governed by another, third variety of control, which he labels *empathic control*. Whereas voluntary control targets actions and evaluative control targets mental attitudes, empathic control, according to Shoemaker, instead targets *reasonish regard*.

What is reasonish regard? When facts about another agent and their interests "properly appear to me as putative reasons, I have reasonish regard for them; when they don't, I don't" (Shoemaker, 2022: 98). For example, if I play loud music late into the night, and the fact that my neighbor will be kept awake by the music doesn't even feature in my deliberations about what to do, then a fact that *should* appear as a reason in my deliberations doesn't, and for that I'm blameworthy.

Shoemaker takes reasonish regard to be a quasi-perceptual mental stance, since he takes it to be passive and therefore unlike either actions or attitudes. Reasonish regard is, rather, a state of *empathy*. It's to take another agent's perspective in a certain sort of way, *as mattering*. Empathy is a form of control, in turn, because a capacity for empathy can be understood in terms of the workings of a relevant *mechanism*, just as in the cases of voluntary and evaluative control.

> To say that one has "empathic control" over whether one has reasonish regard for someone is just to say that one has a normal empathic mechanism wherein the demanded perceptual stance is caused or brought about by robust perspective-taking with that person. So were I to think about how loud and irritating my music must be from the perspective of my neighbor, how she values work and is livid over this din, I should come to perceive those facts from my own perspective as at least putative reasons to turn it down. (2022: 105)

Shoemaker's conclusion is conditional: *if* control is needed for responsibility, then we must make room for empathic control too. A third variety of control might underpin human moral responsibility.

Turning to legal responsibility, Katrina Sifferd and Anneli Jefferson defend a "hybrid" view of the justification for legal punishment. They maintain that while moral desert may be a necessary condition for blame and punishment, it's not sufficient, and "some further instrumental good such as moral development or social order needs to be met" (Sifferd & Jefferson, 2022: 123). Sifferd and Jefferson apply this framework to cases of "reckless rape" and how the law deals with and ought to deal with them. They begin with the legal requirements for the least serious cases of rape, so-called "simple rape." According to the Model Penal Code (MPC):

> A defendant may be found guilty of sexual assault without consent if he causes another person to submit to or perform an act of sexual penetration or oral sex and (a) the other person does not consent to that act; and (b) the actor is aware of, yet recklessly disregards, the risk that the other person does not consent to that act. (MPC 213.6)

As the authors note, these requirements assume that the core wrongdoing in cases of rape consists in engaging in non-consensual sex in a culpable mental state, rather than in the use of physical force or its threat (which might be additionally blameworthy aspects of a given case of rape).

Regarding the culpable mental state of *reckless disregard*, which condition (b) tries to pick out, the authors outline how most interpretations of this condition require that the offender in a case of rape be *consciously* aware of a substantial risk of non-consent. Here, non-consent is characterized in terms of behavior that might be taken to communicate a lack of willingness on the part of the victim. On this view, *negligent* rapists – i.e., those who satisfy the act requirement of condition (a) yet who are not *consciously* aware of a substantial risk of non-consent according to condition (b) – do not count as blameworthy for rape. The authors disagree.

On Sifferd and Jefferson's interpretation of the recklessness condition for simple rape, "negligent rapists are morally blameworthy, but not to a degree where a criminal conviction – particularly conviction of a felony – is a proportionate response" (2022: 129). As a result, "in general, people are morally blameworthy for simple rape in a much wider range of cases than are stipulated in the new MPC Sexual Assault provisions" (2022: 130). On Sifferd and Jefferson's hybrid view of the justification of punishment, blame and punishment of reckless rapists is justified on instrumental grounds, i.e., in terms of its effects.

These effects need not be on the offender themselves but may work by communicating something to the moral community about what counts as acceptable behavior. As they note, "there is... evidence that the law is successful both in communicating and shaping norms, for example in the context of changes in gay marriage law" (2022: 134). Sifferd and Jefferson think that unless we take this function of the law seriously by penalizing cases of reckless rape, we run the risk of implicitly endorsing such behaviors.

Do reckless rapists *deserve* punishment? Or does the instrumental justification for punishment that Sifferd and Jefferson rely upon risk our *wronging* reckless rapists by treating them as mere instruments to communicate things to our wider moral communities? Sifferd and Jefferson say this:

> Persons who meet the MPC requirements for negligence – persons who grossly deviate from a standard of care that a reasonable person would observe during sex – are morally culpable when they rape: however, we think the level of culpability exhibited by negligent rapists is not sufficient to justify a felony conviction. In theory, we are open to the possibility of negligent rapists being found guilty of a misdemeanor, where this conviction does not entail the possibility of custodial sanctions. (2022: 139)

Yet they immediately add that "the positive instrumental effects and the very low culpability taken together may not outweigh the harm done to the offender sufficiently to justify punishing negligent rape as a misdemeanor" (2022: 139–140). If so, some other form of punishment or sanction would, they imply, have to be found, which would be proportionate to these offences.

Sifferd and Jefferson think the objection that, "we are... instrumentalizing current reckless rapists for the sake of societal progress" (2022: 140) may be unavoidable. "This may be troubling," they write, "but it is not clear that it's avoidable if we take norm expressive and instrumental aims to be part of what justifies punishment" (2022: 22). We need, they think, to hold reckless rapists responsible so as to communicate and enforce norms about seeking sexual consent.

In "Punishment and desert," Gregg Caruso argues that retributive punishment is not justified, on theoretical and practical grounds. Caruso starts by recapitulating his previously published work defending *free-will skepticism* – the view that no one has the sort of free will that would justify blame or retributive punishment. There are two versions of this argument. The stronger version claims that the standard positive positions on free will – compatibilism, event-

causal libertarianism, and agent-causal libertarianism – each face sufficient objections that we should abandon them and embrace free-will skepticism instead. This *metaphysical* argument should be recognizable to anyone familiar with free-will debates. The weaker version is *epistemic* and hinges on the idea that there is (or should be) a high epistemic bar for justifying harming people by punishment, and the standard positive positions on free will fall far short of this high bar. Moreover,

> Unlike the Skeptical Argument... the Epistemic Argument does *not* require the refutation of libertarian and compatibilist accounts of free will. Instead, it simply needs to raise sufficient doubt that they succeed. (Caruso, 2022: 153–154)

So much for the *theoretical* objections to retributive punishment. Even more interestingly, Caruso also outlines a number of *practical* objections. First, his *misalignment argument* argues that criminal law is unable to properly distribute punishment according to perpetrators' desert, a key claim of retributivist defenses of punishment, "because criminal law is not properly designed to account for all the various factors that affect blameworthiness, and as a result the moral criteria of blameworthiness are often misaligned with the legal criteria of guilt" (2022: 158).

For example, Caruso points out that many theorists of moral responsibility maintain that mental illness often mitigates blameworthiness, but the law does not reflect such mitigating circumstances (Caruso notes that the legal *insanity defense* is technical and very limited in scope).

But even if we were to fix the legal system by expanding the scope of mitigating circumstances, such that the criteria for punishment more closely tracked those of moral blameworthiness, Caruso argues that the instruments of the state are rarely in an epistemic position to know what an agent deserves, because, for example, of entrenched implicit racial bias, among other things. As a result, the state is unable to distribute punishment according to desert, Caruso maintains.

The upshot is a dilemma. Either retributivist defenders of punishment resist broadening the scope of mitigating circumstances in criminal law or they do not. If they do, they confront the *poor epistemic position* argument, i.e., the claim that the state is almost never in an epistemic position to know what an agent deserves. If they don't, then they confront the *misalignment argument.*

Thus, retributivism is bound to result in injustice, Caruso maintains, even judged on its own standards.

4. Science and agency

As I mentioned earlier, on a thoroughly naturalistic approach to free agency, two obvious and related questions arise. First, how and why, evolutionarily, did we acquire the abilities that underpin our seemingly free agency. Second, how and why have we acquired the concept of free agency – and various adjacent concepts – both in our developmental learning histories and evolutionarily.

In the first paper in Section 4, "Children's developing beliefs about agency and free will in an increasingly technological world," Teresa Flanagan and Tamar Kushnir tackle the latter question. In particular, they examine how children's developing concepts of agency and free will are applied to artificial agents, such as robots. As the authors note, most previous research on children's developing concepts has measured how they are applied to humans. As a result, it has failed to address whether (or when) children might apply these concepts to *non-human* agents. Since today's children frequently and increasingly interact with non-human, robotic agents, it behooves us, the authors maintain, to answer this question. To this end, Flanagan and Kushnir survey extant work on the topic and present the findings of a number of their own studies.

In brief, the authors' studies confirm and expand the findings of earlier work in showing that young children tend to treat robots as agents on a par with humans and as social and moral partners. This contrasts with adults' tendency to treat robots *not* as free agents but merely as helpful tools. However, children's tendency to attribute free agency to robots changes with age, with younger children being more likely to treat robots as acting freely and older children treating them more as tools. Interestingly, the authors point to the fact that these findings are cross-sectional rather than longitudinal, and so they raise the interesting hypothesis that the increasing exposure of today's children to ever more sophisticated robot agents may well result in today's younger children retaining a stronger tendency to attribute free agency to robots as they grow older and into adulthood. The upshot? Stay tuned! Exposure to robots may be changing how we, as humans, apply our concepts of agency, free agency, and social and moral partnership.

Although it's certainly not the authors' intention in their paper, it's easy to see how Dennis Papadopoulos and Kristin Andrews' contribution, "How mindshaping and social maintenance can support shared intentions in great

apes," can be read as a contribution to answering the other question I posed earlier, about how and why, evolutionarily, we acquired the abilities that might underpin our free agency. Specifically, Papadopoulos and Andrews tackle the question of whether shared intentions exist in non-human animals – here, in other great apes. For me, this kind of research is *essential* reading for naturalistic philosophers working on agency and moral responsibility. In particular, given the increasing popularity of views such as Vargas's "Agency Cultivation Model" of our moral responsibility practices, according to which such practices are not only best explained but also justified by how they promote and sustain particular forms of agency in our societies (e.g., Vargas, 2013), it's not only advisable but *required* that we examine how related phenomena might operate in our closest evolutionary cousins.

Papadopoulos and Andrews maintain that other animals, and especially great apes, have *shared intentions* – i.e., individuals do things together in a group rather than acting entirely independently. Shared intentions have two important features – i.e., *joint commitment* and *standing to rebuke*. Papadopoulos and Andrews argue that great apes exhibit both features, contra skeptics who maintain that the features (and thus shared intentions) are unique to humans.

More particularly, a joint commitment requires that two conditions be met: "(i) continuing a joint activity until goals are obtained for all involved, (ii) preferential sharing of rewards with collaborators" (2022: 207). In turn, shared intentions entail a standing to rebuke, "which allows one to predict that their partner would be in a position to protest a deviation from the joint project" (2022: 207). Those skeptics who maintain that shared intentions are unique to humans cite empirical evidence as showing that non-human animals lack these features. Papadopoulos and Andrews show that this evidence doesn't support the skeptical view in anything like the way skeptics think, and moreover there are more plausible interpretations of the evidence that instead support the claim that great apes in fact have shared intentions. I won't review all of the interesting details of Papadopoulos and Andrews' argument. Instead, I leave their investigation to you.

The upshot, however, is this:

> With a broader, and more empirically adequate account of shared intentionality, we can more clearly see a route forward for examining the degree to which other species, including chimpanzees, work together as shared agents. Indeed, we think that an investigation into nonhuman shared intentionality must go hand in hand with the investigation into nonhuman social norms. This development can

> also inform our understanding of the breadth of capacities involved in human social norms and shared intentionality. (2022: 219)

This editor, doffing his editor's hat for a moment to don his philosopher's one, couldn't agree more.

5. Artificial moral agency

In Section 5, on "Artificial moral agency," Marcus Arvan grapples with the question of how we ought to think about the varieties of *artificial* moral agents that we might already be on our way to building. In his paper, "Varieties of artificial moral agency and the new control problem," Arvan first outlines the *control problem*, i.e., the problem of ensuring that humans maintain control over artificial agents, especially when they might pose an existential threat to us. Arvan also outlines the related *alignment problem*, which concerns how to ensure that such agents' goals remain aligned with our own (which presumably helps to address the control problem).

Next, Arvan identifies three potential categories of artificial moral agent: *inhuman* artificial agents, *better-human* agents, and *human-like* agents. In unpacking the details of each of these categories, Arvan maintains that each category raises *distinct* control and alignment problems. Worse, these problems can't be solved for the first category, he thinks – i.e., for inhuman artificial agents. And regarding better-human agents, Arvan thinks that such agents would likely only amplify the various moral errors to which humans are already prone. So, these agents would also be likely to veer out of our control and become misaligned with our goals and values.

The third category – the human-like agents – would probably replicate human moral errors too, Arvan thinks. But much worse than that, the possibility of such agents gives rise to what Arvan calls the *new control problem*. In short, if these artificial agents were indeed sufficiently human-like in being intelligent and conscious, or in having interests and wills of their own, they would presumably be entitled to the same sorts of rights and freedoms as us. The new control problem consists in figuring out the morally appropriate ways in which humans and human-like agents would be permitted to control each other. In short, any respect in which human-like agents might have superiority over us – for example, either physically or cognitively – would, according to Arvan, likely drive a tendency for these agents to exert morally unjustifiable forms of dominance over us. Yet attempts by us to limit these agents' abilities would also likely result in

our harming them in morally unjustifiable ways. We would face a new wrinkle, putting it mildly, in how social justice should work in a society comprising both human *and* artificial agents.

*

With that, I'll wrap up this brief Introduction. The papers that follow are more than capable of speaking for themselves, and I hope I have not done them any injustice in my brief synopses of them.

To close, let me express my gratitude, first, to all of the authors whose work appears herein. You were all gracious enough to accept my invitation to contribute and were enormously patient with, and forgiving of, this novice editor. In turn, I would like to thank those anonymous individuals – you know who you are! – who kindly volunteered their time to review these articles. Reviewing work is often thankless. Let that not be the case here. My deep thanks to you all. Lastly, my thanks to those at *Humana Mente* for entrusting an issue of their esteemed journal to my hands. What a wild thing to do! I hope I haven't let you down. I know the authors haven't.

REFERENCES

Argetsinger, H., M. Vargas. (2022). What's the Relationship Between the Theory and Practice of Moral Responsibility? *Humana.Mente Journal of Philosophical Studies*, 42, 29-62.

Caruso, G. (2015). Free Will Eliminativism: Reference, Error, and Phenomenology, *Philosophical Studies*, 172(110): 2823–2833.

Caruso, G. D. (2022). Punishment and Desert. *Humana.Mente Journal of Philosophical Studies*, 42, 145-178.

Deery, O., K. Bailey. (2022). The Bias Dilemma: The Ethics of Algorithmic Bias in Natural-Language Processing. *Feminist Philosophical Quarterly*, 8(3–4): Feminism, Social Justice, and Artificial Intelligence, 1–28.

Deery, O., Bedke, M., Nichols, S. (2013). Phenomenal Abilities: Incompatibilism and the Experience of Agency. In D. Shoemaker, ed., *Oxford Studies in Agency and Responsibility, Vol. 1*. Oxford: Oxford University Press, pp. 126–150.

Deery, O. (2015) The Fall from Eden: Why Libertarianism Isn't Justified by Experience. *Australasian Journal of Philosophy*, 93(2), 319–334.

Deery, O. (2021a). Free Actions as a Natural Kind. *Synthese*, 198(1), 823–843.

Deery, O. (2021b). *Naturally Free Actions*. Oxford: Oxford University Press.

Flanagan, T. M., T. Kushnir. (2022). Children's Developing Beliefs About Agency and Free Will in an Increasingly Technological World. *Humana.Mente Journal of Philosophical Studies*, 42, 179-204.

Heller, M. (1996). The Mad Scientist Meets the Robot Cats: Compatibilism, Kinds, and Counterexamples. *Philosophy and Phenomenological Research*, 56(2), 333–337.

Horgan, T. (2015). Injecting the Phenomenology of Free Will into the Free Will Debate. In D. Shoemaker ed., *Oxford Studies in Agency and Responsibility, Vol. 3*. Oxford: Oxford University Press, pp. 34–61.

Horgan, T. (2007). Agentive Phenomenal Intentionality and the Limits of Introspection. *Psyche*, 13, 1–29.

Horgan, T. (2011). The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits. *Humana.Mente Journal of Philosophical Studies*, 15, 77–97.

Horgan, T. (2014). Phenomenal Intentionality and Secondary Qualities: The Quixotic Case of Color. In B. Brogaard, ed., *Does Perception Have Content?* Oxford: Oxford University Press, pp. 329–350.

Horgan, T., M. Timmons. Is Agentive Freedom a Secondary Quality? *Humana.Mente Journal of Philosophical Studies*, 42, 63–87.

McCormick, K. (2022). A Discretionary Case for Preservationism about Free Will. *Humana.Mente Journal of Philosophical Studies*, 42, 1–28.

McKenna, M. (2008) Putting the Lie on the Control Condition for Moral Responsibility. *Philosophical Studies*, 129, 2937.

Nichols, S. (2015). "Free Will and Error. In S. Nichols, *Bound: Essays on Free Will and Responsibility*. Oxford: Oxford University Press, pp. 54–74.

Nichols, S. (2017). The Essence of Mentalistic Agents. *Synthese*, 194, 809–825.

Nichols, S. (Forthcoming). Free Will and Reference. In J. Campbell, ed., *A Companion to Free Will*. Hoboken, NJ: Wiley-Blackwell.

Papadopoulos, D., K. Andrews. (2022). How Mindshaping and Social Maintenance can Support Shared Intentions in Great Apes. *Humana.Mente Journal of Philosophical Studies*, 42, 205–223.

Sifferd, K., A. Jefferson. (2022). Responsibility for Reckless Rape. *Humana.Mente Journal of Philosophical Studies*, 42, 119–143.

Shoemaker, D. (2022). Empathic Control? *Humana.Mente Journal of Philosophical Studies*, 42, 89–118.

Strawson, G. (1986). *Freedom and Belief*. Oxford: Clarendon Press.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. New York: Oxford University Press.

Vargas, M. (Forthcoming). Revisionism. In J. Campbell, ed., *A Companion to Free Will*. Hoboken, NJ: Wiley-Blackwell.