Review of International Studies (2024), **50:** 3, 534–559
doi:10.1017/S0260210524000202

RIS

SPECIAL ISSUE ARTICLE

AI and the future of IR: Disentangling flesh-and-blood, institutional, and synthetic moral agency in world politics

Toni Erskine

Coral Bell School of Asia Pacific Affairs, Australian National University (ANU), Canberra, ACT, Australia
Email: toni.erskine@anu.edu.au

(Received 1 August 2023; revised 2 February 2024; accepted 4 February 2024)

Abstract

Intelligent machines – from automated robots to algorithmic systems – can create images and poetry, steer our preferences, aid decision making, and kill. Our perception of their capacities, relative autonomy, and moral status will profoundly affect not only how we interpret and address practical problems in world politics over the next 50 years but also how we prescribe and evaluate individual and state responses. In this article, I argue that we must analyse this emerging synthetic agency in order to effectively navigate – and theorise – the future of world politics. I begin by outlining the ways that agency has been under-theorised within the discipline of International Relations (IR) and suggest that artificial intelligence (AI) disrupts prevailing conceptions. I then examine how individual human beings and formal organisations – purposive actors with which IR is already familiar – qualify as moral agents, or bearers of duties, and explore what criteria intelligent machines would need to meet to also qualify. After demonstrating that synthetic agents currently lack the 'reflexive autonomy' required for moral agency, I turn to the context of war to illustrate how insights drawn from this comparative analysis counter our tendency to elide different manifestations of moral agency in ways that erode crucial notions of responsibility in world politics.

Keywords: agency; AI; ethics of war; International Relations theory; moral agency; moral responsibility

Over the next 50 years, artificial intelligence (AI) – the evolving capability of machines to imitate aspects of intelligent human behaviour – will continue to unsettle every aspect of world politics. This disruption will include: AI's (unequal) effect on the global workforce, along with exacerbated disparities in the distribution of resources and wealth; governments' enhanced abilities to monitor, control, and judge their citizens, as well as intervene in the politics of other states; and the impact of AI-enabled systems on both the conduct of war and, inevitably, states' decision making on the resort to force. Intelligent machines – from algorithmic systems to automated robots – can already create images and poetry, steer our preferences, aid decision making, and kill.[1] Our perception of the capacities, relative autonomy, and status of these artefacts will profoundly affect not only how we interpret and address practical problems in world politics. It will also affect what we expect of the individual human actors and states that employ them – and how we ethically evaluate *our* acts and omissions. In other words, how we understand these increasingly sophisticated intelligent machines represents both an urgent practical challenge in world politics and a neglected theoretical puzzle for the discipline that seeks to study it.

[1]For the purposes of this article, I will refer to intelligent machines, intelligent artefacts, and AI-enabled entities/systems interchangeably.

At this point in time, when we consider the status of those AI-enabled entities that are exerting unexpected effects on international politics, one observation should be uncontroversial. Namely, whether we are talking about seemingly ubiquitous language-generative models such as ChatGPT, or the Super aEgis II armed robotic sentry that has been tested in the demilitarised zone between North and South Korea, these AI-enabled systems are not moral agents, or duty-bearers in their own right. They are not bodies that we can reasonably expect to 'do the right thing' or bear the burden of blame in the wake of harm or wrongdoing. Simply, current incarnations of intelligent machines lack the specific capacities that would allow them to qualify as moral agents. Rather, they are the *tools* of the individual human and state actors that employ them. As straightforward as this may seem, we need to be able to defend and unpack this assertion – and reassess it as technologies rapidly develop. Indeed, as science fiction appears to merge with reality, and AI-enabled entities seem to both slouch towards autonomy and become 'more human', the question of where moral responsibility lies for specific acts and outcomes that involve sophisticated forms of AI will be posed with increasing urgency – and uncertainty. Understanding what it would take for these entities to qualify as moral agents in their own right – and, by extension, why they do not currently warrant this label – becomes extremely important.

The discipline of International Relations (IR) should be well equipped to contemplate the parameters of a sophisticated new *synthetic* agent. After all, it already acknowledges very different embodiments of the purposive actor in world politics. Unfortunately, it also has a track record of spectacularly under-theorising agency. Prominent positions within IR wilfully blur important distinctions between the respective agency of individual human beings and states, for example, and historically all but ignore key questions of moral agency. As such, IR lacks the conceptual tools needed to interrogate the status of intelligent machines in world politics, situate them in relation to existing moral agents, and guard against already-apparent conceptual entanglements that undermine our ability to accurately describe, prescribe, and evaluate actions in relation to them. In short, when it comes to both understanding and ethically evaluating purposive action in world politics, the puzzle of AI reveals that IR theory is not (yet) fit for 21st-century purpose.

One way of addressing this shortcoming is to re-examine how we understand the sophisticated purposive actors with which we are already familiar in world politics, and about which we make such bold assumptions in IR. What are the defining features that make not only individual human actors but also states and other corporate entities possible loci of moral responsibility? What can this tell us, in turn, about what it would take for those intelligent artefacts that are already fundamentally affecting the fabric of world politics to become more than just our tools and also qualify as moral agents? (Or to pose the question differently: What does this tell us about why they do not presently qualify?) Moreover, and more speculatively, if intelligent machines were to qualify as moral agents at some point in the future, how would they differ from (and resemble) the flesh-and-blood and corporate conceptions that IR currently, at least implicitly, acknowledges? And, finally, what are the practical risks of sidestepping these considerations?

In attempting to answer these questions, this article will be divided into four sections. The first section will briefly set out the ways that agency, and specifically moral agency, have traditionally been under-theorised within the discipline of IR. It will also suggest that the advent of AI offers both an alternative to prevailing conceptions and an opportunity to imbue them with greater nuance. The second section will revisit the two categories of moral agent in world politics that we currently recognise: individual human actors, generally considered to be archetypal moral agents, and corporate bodies that qualify as what we might call 'institutional moral agents'. Inspired by the work of Onora O'Neill, I will propose key points of convergence and contrast between these two types of moral agent. The third section will draw on this comparative analysis – and accompanying case for a non-human variation on moral agency – to consider what features intelligent artefacts would have to possess if they could ever reasonably be expected to discharge duties and be considered

culpable.[2] In short, I will sketch a model of what I will label 'synthetic moral agency'. The final section will illustrate why this analysis and resulting typology matter in practice by turning to our perceptions and expectations of intelligent artefacts in the context of organised violence. After introducing the category of 'moral agents of restraint', and reiterating that machines cannot (yet) qualify as such, I will suggest that our failure to examine the status of AI-enabled entities invites two worrying instances of 'misplaced responsibility' in war. These perilous misattributions result from the unexamined assumptions that either we know what as-yet-hypothetical synthetic moral agents would look like (and they would look like *us*), or, more immediately, *they already exist*. Analysing and countering such assumptions is one of the most significant theoretical and practical challenges that IR will face over the next five decades.

Before proceeding, two points of clarification on the aims of this article are in order, one related to its theoretical engagement with IR and the other with respect to its practical focus on world politics. First, while this article makes a contribution to normative IR theory, where questions of moral agency are core concerns, it also seeks to speak to IR more broadly. It demonstrates that extending our current conception of purposive action in world politics is – and will increasingly be – crucial to IR's explanatory and normative endeavours. Simply, puzzles posed by the evolution of AI reveal that a more nuanced understanding of agency, and specifically moral agency, is fundamental to the future of IR theory if it is to remain relevant to pressing 21st-century concerns. Second, although this article uncovers a significant practical risk that accompanies the global proliferation of AI, the aim here is not to join the current cautionary chorus lamenting an anticipated existential threat posed by AI's future iterations. Its focus is much nearer. Frequently articulated fears about advanced AI systems of the future – systems that would possess the potential for catastrophic harm, yet be beyond the power of individuals, states, or humanity as a whole to control – eclipse a more mundane, more insidious, and more immediate hazard. This article suggests that a neglected danger of already-existing AI-enabled tools is that they change how *we* (as citizens, scholars, soldiers, or states) deliberate, how *we* act, and how *we* view ourselves as responsible agents in world politics. This has potentially profound ethical, political, and even geopolitical implications – and is directly related to how we understand the capacities and concomitant status of intelligent machines.

(Moral) agency, AI, and the limits of IR theory

IR's tendency to make bold assumptions about the capacities of certain corporate entities in world politics – assumptions with implications not only for their ontological but also their moral status – provides a fascinating starting point for questioning the moral agency of AI-enabled entities. However, IR theory has historically under-theorised agency.[3] Notably, it is hampered by four limitations that are particularly relevant to the current analysis. These limitations prompt questions anew when we introduce the possibility of intelligent machines as an emerging category of purposive actor in world politics.

First, prominent theoretical approaches within IR have embraced a notoriously narrow conception of agency fashioned on an idealised account of the individual human being as an independent, unitary, and perfectly rational actor. This conception has been applied to the state through methodologically motivated, rudimentary analogies, which (in the language of Richard Ashley's cutting critique of the 'poverty' of neo-realism) take states to be 'the *living individuals* of international life'[4]

[2]I will treat moral responsibilities, duties, and obligations interchangeably for the purposes of this article. However, I acknowledge that various conceptual distinctions can be made. See, for example, Joel Feinberg, *Doing & Deserving: Essays in the Theory of Responsibility* (Princeton, NJ: Princeton University Press, 1970), pp. 132–42, and Robert E. Goodin, *Utilitarianism as a Public Philosophy* (Cambridge: Cambridge University Press, 1995), pp. 81–7.

[3]This is not a new charge. See similar statements in Arnold Wolfers, 'The actors in international politics', in *Discord and Collaboration: Essays in International Politics* (Baltimore, MD: Johns Hopkins University Press, 1962), pp. 3–24; Alexander Wendt, *Social Theory of International Politics* (Cambridge: Cambridge University Press, 1999); and Colin Wight, *Agents, Structures and International Relations: Politics as Ontology* (Cambridge: Cambridge University Press, 2006).

[4]Richard Ashley, 'The poverty of neorealism', *International Organization*, 38:2 (1984), pp. 225–86 (p. 239, n. 34).

and thereby elevate a model of agency that accurately represents neither. Distinctions between different types of purposive actor in world politics are consequently overlooked or conveniently ignored in the name of elegant theorising. I am not suggesting that there are no valid points of comparison between the agency of states and individual human beings. (To the contrary, I will argue that important commonalities exist – and, in fact, have been understated in one significant respect.) Rather, the problem is that an airbrushed portrait of individual human agency is accepted as the state-as-agent image underlying many IR theories, obscuring significant details of each and blurring the lines between these categories of purposive actor.

Second, and relatedly, the proposed corporate agency of the state has been generally poorly defended and ill-defined. While classical realist, neorealist, neoliberal institutionalist, and some constructivist approaches have long relied on a bold conception of the state as an agent in its own right, the same positions have not been concerned to defend this status – or explore what makes it unique. It may well be that particular idealised images of agency within the discipline would be difficult to defend and distinguish in this way. Indeed, some prominent theorists who rely on a conception of the state as agent quietly concede that this claim should not be taken too seriously.[5] Yet IR's pervasive state-as-agent assumption need not be conceived merely metaphorically. Compelling defences *are* possible, as demonstrated within IR through path-breaking work by Alexander Wendt, culminating in his 1999 work, *Social Theory of International Politics*.[6] Notably, however, Wendt's iconic claim that 'states are people too'[7] anticipated a regressive tendency in contemporary IR theory among positions that seek to defend the agency of the state. Namely, even while acknowledging the distinct corporate nature of state agency, they revive anthropomorphising moves – seeming to link the adequacy of a defence of state agency to how closely the state can be shown to mirror flesh-and-blood individuals in every respect, including, for example, the possession of consciousness.[8] Indeed, in subsequent writing, Wendt worries that his theory of 'the state as person' allows only 'an impoverished and truncated' kind of person – an '"artificial" person' rather than a '"natural" one' – in the absence of an account of 'collective consciousness'.[9]

Third, IR has lacked consistency in identifying and defending those bodies in world politics that qualify as purposive actors in their own right. Many of the same theoretical approaches that blindly accepted the agency of the state ignored, or explicitly rejected, the agency of other bodies with arguably comparable capacities, including, for example, intergovernmental organisations (IGOs). IGOs have often been conceived as 'instruments' of states and 'structures' within which states pursue their interests instead of as purposive actors. As Michael Barnett and Martha Finnemore incisively observed in 1999, neorealists and neoliberal institutionalists share the conviction that IGOs 'have no ontological independence'.[10] Notably, not all IR theorists embraced this stance and, indeed, an increasing number have since recognised IGOs as independent actors rather than mere

[5]For example, in *Theory of International Politics* (Long Grove, IL: Waveland Press, 1979), p. 119, Kenneth N. Waltz acknowledges that 'we can freely admit that states are in fact not unitary, purposive actors'. See also the discussion in Ashley, 'The poverty of neorealism', pp. 238–9.

[6]Wendt, *Social Theory*.

[7]Wendt, *Social Theory*, p. 194.

[8]See, for example, Alexander Wendt's examination of collective consciousness and concession that perhaps 'states cannot be persons in the full, conscious sense' in 'The state as person in international theory', *Review of International Studies*, 30:2 (2004), pp. 289–316 (p. 314); and, conversely, Adam Lerner's intriguing, and arguably ethically problematic, attempt to defend state consciousness in 'What's it like to be a state? An argument for state consciousness', *International Theory*, 13:2 (2021), pp. 260–86.

[9]Wendt, 'The state as person', p. 313. Wendt's explanation for 'tak[ing] "actor" and "person" to be synonymous' is revealing. Namely, he notes that 'the attributes routinely applied in IR to state actors are those of persons'. This provides a conceptual straitjacket of sorts – especially as the attribution of (ostensibly) human characteristics to states in IR theory is problematic in a number of respects, as addressed above. Wendt adds that 'whether or not there are other kinds of actors I shall not address here'. Arguably, his starting point discourages such an investigation. See Wendt, 'The state as person', p. 289, n. 1.

[10]Michael Barnett and Martha Finnemore, 'The politics, power, and pathologies of international organizations', *International Organization*, 53:4 (1999), pp. 699–732 (p. 704).

instruments to serve state interests.[11] Nevertheless, many take this position for granted, with few attempting an explicit defence of how (and when) IGOs become corporate agents in a way not ultimately reducible to their constitutive parts.[12]

Fourth, with the exception of some recent work in normative IR theory (which will be addressed below), IR has failed to take seriously the ethical implications of its commitment to certain corporate entities as sophisticated purposive actors. Whether taking for granted an analogy between the individual human actor and the state, providing a sustained, if still anthropocentric, account of the state as agent, or simply assuming a conception of corporate agency (that may or may not include intergovernmental and non-state actors), IR theories have traditionally stopped short of taking this confident commitment to its logical conclusion. The assumptions of corporate agency that are widely made (and sometime defended) in much IR theorising valuably gesture towards a specifically *moral* agency.[13] This is profoundly important. The discipline's prevalent claims that certain organisations are agents with interests, aims, and sophisticated decision-making capacities necessarily have implications for what we can expect of these bodies, and whether (and when) we can hold them to account for particular acts and omissions. In other words, they matter in the context of consequential moral responsibility judgements that are regularly made in world politics. Let me explain.

Attributions of moral responsibility are powerful means of guiding and censuring conduct in world politics. They can take the form of either prospective judgements, articulated in the language of duty and obligation, or (intimately linked) retrospective evaluations, voiced most frequently in international politics through assertions of blame and accountability. To be coherent, each attribution must be directed towards a *moral agent*. To avoid confusion, it is important to note that this label does not designate an agent that is necessarily good, just, or otherwise ethically estimable. (That is not how the qualifier 'moral' is being used here.) Rather, moral agents are agents that are *capable of specific types of reasoning*. As such, we can justifiably have certain expectations of them. Moral agents are sophisticated purposive actors that possess capacities for deliberation, for understanding and reflecting on their actions and the probable outcomes of their actions, for evaluating their reasons for adopting a particular course of action, and for acting on the basis of this deliberation and self-reflection.[14] Because they possess these capacities, we can reasonably expect them to understand what are deemed to be moral requirements, and (if other conditions are met) to

[11]See, for example, Gayl D. Ness and Steven R. Brechin, 'Bridging the gap: International organizations as organizations', *International Organization*, 42:2 (1988), pp. 245–73; Duncan Snidal, 'Political economy and international institutions', *International Review of Law and Economics*, 16:1 (1996), pp. 121–37; Martha Finnemore, *National Interests in International Society* (Ithaca, NY: Cornell University Press, 1996); Michael Barnett and Martha Finnemore, *Rules for the World: International Organizations in Global Politics* (Ithaca, NY: Cornell University Press, 2004); Kenneth W. Abbott and Duncan Snidal, 'Why states act through formal international organizations', *The Journal of Conflict Resolution*, 42:1 (1998), pp. 3–32; Darren G. Hawkins, David A. Lake, Daniel L. Nielson, and Michael J. Tierney, *Delegation and Agency in International Organizations* (Cambridge: Cambridge University Press, 2006); and Michael Zürn, Martin Binder, and Matthias Ecker-Ehrhardt, 'International authority and its politicization', *International Theory*, 4:1 (2012), pp. 69–106.

[12]Exceptions include Toni Erskine, '"Blood on the UN's hands"? Assigning duties and apportioning blame to an intergovernmental organisation', *Global Society*, 18:1 (2004), pp. 21–42; and, more recently, Matthias Hoffenberth, 'Get your act(ors) together! Theorising agency in global governance', *International Studies Review*, 21:1 (2019), pp. 127–45; Thomas Gehring and Kevin Urbanski, 'Member-dominated international organizations as actors: A bottom-up theory of corporate agency', *International Theory*, 15:1 (2023), pp. 129–53; and Thomas Gehring, 'International organizations as group actors: How institutional procedures create organizational independence without delegation to institutional agents', *Historical Social Research*, 48:3 (2023), pp. 94–124.

[13]See Toni Erskine, 'Locating responsibility: The problem of moral agency in international relations', in Christian Reus-Smit and Duncan Snidal (eds), *The Oxford Handbook of International Relations* (Oxford: Oxford University Press, 2008), pp. 699–707.

[14]There is a conceptual distinction to be made between agency and moral agency. While bodies are agents if they are capable of some degree of purposive action, the moral agents to which any coherent judgement of moral responsibility must be directed necessarily clear a higher bar by also possessing the capacities highlighted here. For example, when we turn to the individual human case, young children are agents (with moral standing), but not generally considered to be moral agents. (See Toni Erskine, 'Making sense of "responsibility" in international relations: Key questions and concepts', in *Can Institutions Have*

act in such a way as to conform to them. In other words, these capacities render them liable to the ascription of duties and the apportioning of moral praise and blame in the context of specific actions. Significantly, positions across a range of IR theories not only make explicit assumptions about certain corporate entities in world politics being purposive actors in their own right but also imbue them with the capacities required for specifically moral agency. This is a crucial move. Yet mainstream IR theory, methodologically predisposed to eschew ethical analyses, has failed to acknowledge the implications of this move – i.e. that bodies with such impressive capacities can be held morally responsible for their actions.

Of course, other considerations also come into compelling judgements of moral responsibility in world politics. Such judgements appeal to institutionalised norms – or evidence of the shared understandings that prefigure them – regarding what constitutes conduct that is right or wrong, just or unjust, required or prohibited.[15] Moreover, to be subject either to the assignment of duties or to the apportioning of blame in the context of specific acts or omissions, the moral agent in question must possess both the specific competencies to perform (or to have performed) the requisite action[16] and must also enjoy the freedom to act (or to have acted), unimpeded by constraining structures, a lack of resources or power, intervening agents, or forces beyond the agent's control.[17] Crucially, though, before determining whether duties can be assigned or blame apportioned in particular circumstances, we need to know that we are talking about a moral agent. This status matters for both IR's explanatory and normative pursuits.

So where do IR's limits and possibilities in theorising agency leave us in our attempt to decipher the moral status of emerging AI-enabled entities in world politics? Each shortcoming is amplified when we address it in light of speculation about the ontological and moral status of intelligent machines. First, not only does AI come closer to the narrow model of purely rational actor that much IR theory has embraced than the actors that the model has claimed to represent, but it also reveals the need for a more nuanced and varied understanding of agency that would allow for important differences that we intuitively know exist. Second, attention to intelligent machines challenges common distinctions between 'artificial' and 'natural' agents and suggests that if sophisticated expressions of agency can take multiple forms, then we need to understand their defining and distinguishing features, without relying on a single, anthropocentric mould. Third, consideration of AI-enabled entities reinforces the coherence and commonality across different forms of corporate agent in world politics and reminds us of the ways in which they are different from individual human actors. While corporate entities as a class of agent in world politics share some capacities with their flesh-and-blood counterparts, they also reveal commonalities with intelligent artefacts that do not extend to individual human agents. In short, they boast a distinct set of defining characteristics. Finally, IR's core assumption that purposive actors with sophisticated capacities exist beyond individual human beings provides a powerful provocation when it comes to questioning the moral status of the increasingly intelligent machines in our midst.

In response, the following two sections will offer a preliminary typology of moral agency in world politics. Three models will be explored, compared, and contrasted. Two are ostensibly well

Responsibilities? Institutional Moral Agency and International Relations (New York: Palgrave Macmillan, 2003), pp. 1–16 [p. 6].) The same is true of non-human animals. The focus of what follows is on different embodiments of specifically moral agency.

[15]I am not making an argument here about the source or derivation of such moral norms in international politics. Accounts of how particular moral responsibilities are grounded and justified are multiple and contested. I take it as given that there is nevertheless broad agreement on some moral responsibilities – across borders and values systems – and that these exert considerable influence even when not universally adhered to. My focus is, instead, on the bodies that can reasonably be expected to respond to *what we understand to be* ethical imperatives.

[16]My point is simply that capacities *in addition* to the threshold capacities needed to qualify as a moral agent are required to be liable to the ascription of particular duties.

[17]Put differently, to reasonably expect a moral agent to discharge a duty, the agent must enjoy the external conditions necessary to perform the requisite actions. To be an appropriate object of blame for some failure (or, indeed, praise for some estimable act or outcome), the moral agent must have been able to do otherwise in the context of the act or omission being evaluated.

known within IR, yet remain under-theorised; the third has been completely neglected within the discipline and is, as yet, only a future possibility in practice.

Revisiting flesh-and-blood and institutional moral agents in world politics

In beginning to address the status of intelligent machines, I will turn to our two most obvious candidates for moral agency in world politics: individual human beings and formal organisations. My reason for considering them together is twofold. First, I want to highlight the increasingly accepted position that human beings do not exhaust the category of moral agent. Second, it is important to make the straightforward, yet often overlooked, point that those entities that occupy different categories of moral agency have distinct defining features, unique capacities, and particular limitations – despite sharing the basic threshold capacities that allow them to qualify as duty bearers. These points will set the stage for the arguments that follow in the third and fourth sections. Not only does reflecting on how we understand flesh-and-blood and corporate moral agents tell us something about the potential moral agency of intelligent artefacts, but understanding the differences between all three (existing and potential) types of moral agent is important when we seek to assign responsibilities and apportion blame in practice.

Meeting the threshold capacities for moral agency

Individual human beings generally provide the model for our understanding of moral agency. Yet we can disaggregate what it means to be a moral agent from what it means to be human. Not all human beings qualify as moral agents. Those who do not qualify – the very young and some with certain intellectual and developmental disabilities or mental illnesses, for example – are considered no less human. Rather, we do not have the same expectations of them. We do not consider them to be bearers of duties or blame them for particular acts and omissions. Being human is, uncontroversially, not sufficient to qualify as a moral agent. More controversially, one might argue that it is *not necessary*.

As a specific category of moral agent, we are living beings. We are defined by our vulnerability, mortality, sociality and interdependence, incomplete knowledge, and a rationality that is variable, unpredictable, and coloured by emotions. We are unitary beings, but hardly consistent or unchanging over time. Our frailties and fallibilities are significant when considering the possibility of other categories of moral agency. Here one might look to the work of Onora O'Neill. As part of a powerful argument that individual human beings are not the only agents to which ethical reasoning is accessible, O'Neill makes the important move of acknowledging our radically imperfect capacities for deliberation and action. She observes that we take individual human beings to be models of moral agency despite what she pointedly describes as our *limited* rationality, understanding, powers of action, independence, and unity.[18] That most adult human beings are understood to have the requisite capacities to qualify as moral agents tells us something important about our implicitly accepted modest threshold for moral agency – even if IR tends to pay homage to a markedly different ideal rational actor.

The capacities necessary for moral agency proposed in the first section – sophisticated, integrated capacities for deliberation, reflexivity, and action – are not limited to individual human actors. Mainstream theories within IR have travelled a considerable distance towards acknowledging this. Indeed, they have been at the forefront of identifying certain collectivities – particularly states – as agents in their own right with impressive capacities for decision making and action. While they have stopped short of taking this account to its logical conclusion and recognising these collectivities as specifically *moral* agents, theirs is an important preliminary move. This preliminary move is taken further by philosophical arguments that explicitly defend formal organisations

[18] Onora O'Neill, 'Who can endeavour peace?', *Canadian Journal of Philosophy*, supplementary volume 12 (1986), pp. 41–73 (pp. 53, 54, 62).

as potential bearers of moral responsibilities and interrogate the features that allow them to qualify as such.[19] For example, one might argue that a collectivity with the following characteristics qualifies as an 'institutional moral agent':[20] a corporate identity (or an identity that is more than the sum of identities of its constitutive parts); a decision-making structure that can both commit the group to a policy or course of action that is different from the individual positions of some (or all) of its members and allow it to reflect on and evaluate its reasons for acting; mechanisms by which group decisions can be translated into action; an identity over time; and a conception of itself as a unit. According to this account, institutional moral agents include most states, transnational corporations, non-governmental organisations (NGOs), and, at least transiently, IGOs.[21]

There are compelling reasons for recognising such formal organisations as institutional moral agents. They have capacities to deliberate, to reflect, and to act – and, indeed, to both cause and remedy harm on a scale that is well beyond that of any individual human being. Specifically, these structured institutions have capacities for deliberation that are manifest both in highly developed mechanisms for gathering and analysing information and in decision-making procedures that produce, at the corporate level, something analogous to intentions. Such bodies also demonstrate a capacity for reflexivity, which is evident, for example, as a key component of what might be called 'institutional learning'.[22] Their decision-making structures also allow them to reflect on their conduct, and on the consequences of their previous acts and omissions, evaluate both in light of either external expectations or their internal goals and espoused values, and, as a result, commit to revising (or reinforcing) their own rules, procedures, and organisational culture.[23] Moreover, these formal organisations are able to realise group decisions by coordinating the roles of their constituents and achieving complex levels of integrated action within established frameworks of norms and practices. In the context of some acts and outcomes, ascriptions of duty or blame therefore risk being radically incomplete or misdirected if attached only to individual human beings. Proposed imperatives to engage in, or refrain from, organised violence, as well as condemnations of wars of aggression, provide just such instances. Although individual human actors remain responsible for their own concurrent and complementary contributions, waging war is necessarily a corporate act. If we are unable to describe it as such, our theories lose explanatory power. If we ignore the moral status of corporate actors like states and IGOs, our theories are deprived of normative force.

While still a minority position within mainstream IR (where sophisticated forms of corporate agency are readily assumed but the ethical implications of this assumption are generally eschewed), the recognition of formal organisations as specifically moral agents has become increasingly

[19]For example, Peter A. French, *Collective and Corporate Responsibility* (New York: Columbia University Press, 1984); O'Neill, 'Who can endeavour peace?'; Toni Erskine, 'Assigning responsibilities to institutional moral agents: The case of states and quasi-states', *Ethics & International Affairs*, 15:2 (2001), pp. 67–85; Philip Pettit, 'Responsibility incorporated', *Ethics*, 117:2 (2007), pp. 171–201; Christian List and Philip Pettit, *Group Agency: The Possibility, Design, and Status of Corporate Agents* (Oxford: Oxford University Press, 2011); Stephanie Collins, *Group Duties: Their Existence and Their Implications for Individuals* (Oxford: Oxford University Press, 2019).

[20]See Erskine, *inter alia*, 'Assigning responsibilities'; '"Blood on the UN's hands"?'; 'Coalitions of the willing and responsibilities to protect: Informal associations, enhanced capacities, and shared moral burdens', *Ethics & International Affairs*, 28:1 (2014), pp. 115–45; and 'Intergovernmental organisations and the possibility of institutional learning: Self-reflection and internal reform in the wake of moral failure', *Ethics & International Affairs*, 34:4 (2020), pp. 503–20. The decision-making criterion offered here is inspired by Peter French's account of 'corporate moral personhood'. It is also influenced by Philip Pettit's important work on why certain decision-making structures make group agency possible but is less stringent than Pettit's account. See French, *Collective and Corporate Responsibility*, chapters 3–4, and Philip Pettit, *A Theory of Freedom: From the Psychology to the Politics of Agency* (New York: Oxford University Press, 2001), ch. 5.

[21]IGOs are 'transient' agents and moral agents because they balance intergovernmental structures and deliberative processes with a commitment to member states' sovereignty in ways that can, intermittently, impede their capacity for purposive action at the corporate level. See Erskine, '"Blood on the UN's hands"?', p. 41.

[22]I propose that institutional moral agents are able to learn in a way that is not reducible to learning achieved by their members in Erskine, 'The possibility of institutional learning'.

[23]Erskine, 'The possibility of institutional learning', pp. 508–9.

accepted, and even taken for granted, by explicitly normative approaches within the discipline.[24] This embrace of a tangible alternative to flesh-and-blood moral agents not only opens up important ethical analyses across a range of cases in international politics, but also provides a valuable – and perhaps slightly unsettling – point of departure for exploring the moral status of intelligent machines.

Although individual human beings and formal organisations share the capacities necessary to qualify as moral agents, understanding their respective properties, which produce unique strengths and limitations, is important. This not only helps us to distinguish the requisite features of moral agency from characteristics linked to particular embodiments but is also important when we go on to talk about the moral responsibilities that can reasonably be ascribed to those in each category in contexts such as war. As further points of comparison between flesh-and-blood and institutional moral agents, I will ask where the constituents of each category sit in relations to the common, and often misleading, distinction between 'natural' and 'artificial' agents, and, moreover, whether they qualify as 'moral patients', or bodies to which duties are owed.

'Natural' vs 'artificial' agents

Individual human beings are often described as 'natural' agents, as opposed to 'artificial' agents such as states (or intelligent machines). In the context of her argument that institutional agents have more in common with their individual human counterparts than we may assume, O'Neill suggests that the ostensible divide between 'natural persons' and 'artificial persons' is neither clear-cut nor particularly helpful. She observes that 'when individual agents take on roles, and so acquire distinctive (restructured, extended, diminished) capacities to act and abilities to foresee, they acquire capacities that they would not naturally have had'.[25] Indeed, our social contexts and roles directly affect – both positively and negatively – our access to information and resources, our independence from other agents, and our power to formulate and pursue our own plans. We human beings *are* artificial agents to the extent that our capacities for deliberation and action – and, indeed, our opportunities to actually exercise moral agency in a given situation – are variously enhanced and constrained by the social world around us.

O'Neill's point that the divide between flesh-and-blood and institutional agents is overdrawn when they are described, respectively, as 'natural' and 'artificial' is a valuable one – especially if it encourages us to reconsider the particular features of each. Before doing so, however, it may be useful to pause and note that the modifier 'artificial' may be used to illustrate different things, a point that O'Neill does not address. The description 'artificial' may be understood in (at least) three ways: as socially constituted, in terms of role-defined enhancements and limitations (as O'Neill uses the term to describe individual human agents, above); as counterfeit in contrast to genuine; or as human-made as opposed to biological. Although not directly acknowledged by O'Neill, institutional moral agents are also artificial in the first sense: their behaviour is deeply affected by the systems of social meaning in which they are embedded. They, too, take on roles and are affected by accompanying expectations. They are not, however, artificial in the second sense. As should be clear, I am not presenting the agency or moral agency of corporate entities as a 'fiction' (which

[24]For example, Toni Erskine (ed.), *Can Institutions Have Responsibilities? Collective Moral Agency and International Relations* (London: Palgrave Macmillan, 2003); Mlada Bukovansky, Ian Clark, Robyn Eckersley et al., *Special Responsibilities: Global Problems and American Power* (Cambridge: Cambridge University Press, 2012), pp. 65–6; Neta C. Crawford, *Accountability for Killing: Moral Responsibility for Collateral Damage in America's Post-9/11 Wars* (Oxford: Oxford University Press, 2013), ch. 6; David J. Karp, *Responsibility for Human Rights: Transnational Corporations in Imperfect States* (Cambridge: Cambridge University Press, 2014), pp. 8–11; Sean Fleming, 'Moral agents and legal persons: The ethics and the law of state responsibility', *International Theory*, 9:3 (2017), pp. 466–89 (pp. 468–72); Hannes Hansen-Magnusson and Antje Vetterlein (eds), *The Rise of Responsibility in World Politics* (Cambridge: Cambridge University Press, 2020). As Fleming ('Moral agents', p. 470) observes, 'many works in international political theory now take the idea of corporate moral agency as a basic premise'.

[25]Onora O'Neill, 'Agents, agencies and responsibility', in *Science, Technology and Social Responsibility: Discussion Meeting Held at the Royal Society on Tuesday 16 March 1999* (London: Royal Society, 1999), pp. 13–19 (p. 15).

I take to mean something that is literally and transparently inaccurate, but in a way intended to serve a particular purpose).[26] Rather, I understand formal organisations to be genuine moral agents. Finally, if by 'artificial' one means non-biological, one might note, as O'Neill does, that individual human beings are fundamental components of institutional agents.[27] Yet it is important to qualify that this biological dimension is only part of what constitutes an institutional moral agent. An institutional moral agent also comprises norms, rules, procedures, practices, and organisational culture, which, importantly, serve to frame and channel the decisions and actions of its individual human constituents and allow the organisation itself to behave in ways that cannot adequately be described in terms of the sum of these individual decisions and actions. This combination of flesh-and-blood constituents and a formal structure that coordinates discrete parts so that they work together as a functional whole brings to mind Max Weber's striking image of a '*living machine*', which he invokes to describe 'bureaucratic organisation'.[28] Like flesh-and-blood moral agents, institutional moral agents are emphatically both natural and artificial.

Moral agency and moral patiency

Although both flesh-and-blood and institutional moral agents straddle the natural–artificial divide, they diverge quite radically in another respect. Only flesh-and-blood moral agents are also 'moral patients'. Moral patiency is a philosophical concept that we would do well to introduce into IR discussions, particularly those that seek to understand agency in world politics. Moral patients are objects of moral concern, entities to which duties are owed, entities that have value in themselves. Whereas moral agents are accountable for (at least some of) their acts and omissions, moral patients *count*. In short, they have moral standing. This status can be compellingly grounded in vulnerability to pain and suffering.[29] All individual human beings (whether or not moral agents) are understood to be moral patients, or those to whom we must give moral consideration.[30] Despite their impressive capacities, institutional moral agents are *not* moral patients.[31] They are not sentient. Unlike human beings, they do not have intrinsic value in a way that would make them objects of moral concern. Rather, they have instrumental value. Although formal organisations can be the bearers of moral duties due to their sophisticated, integrated capacities for deliberation, reflexivity, and action, they cannot be (non-instrumental) objects of moral concern.[32] Here the disanalogy between flesh-and-blood and institutional moral agents is particularly acute.

[26] Legal theorists, for example, have traditionally understood the 'personality' of groups such as corporations as useful 'fictions'. See the valuable discussion of this position in Larry May, *The Morality of Groups: Collective Responsibility, Group-Based Harm, and Corporate Rights* (Notre Dame, IN: University of Notre Dame Press, 1987), pp. 11–14.

[27] O'Neill, 'Agents, agencies and responsibility', p. 15.

[28] Max Weber, 'Parliament and government in Germany under a new political order: Towards a political critique of officialdom and the party system', in Ronald Speirs and Peter Lassman (eds), *Weber: Political Writings* (Cambridge: Cambridge University Press, 1994), pp. 130–271 (p. 158), emphasis in original.

[29] Although I take sentience to be the most appropriate property for grounding moral standing, others invoke different properties. For a comprehensive overview, see Agnieszka Jaworska and Julie Tannenbaum, 'The grounds of moral status', in Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 edition), available at: {https://plato.stanford.edu/archives/spr2023/entries/grounds-moral-status/}. Note that moral standing need not entail *equal* moral standing, and, indeed, moral patients with divergent vulnerabilities and interests are commonly understood to make different types of moral claims on us, and to varying degrees.

[30] Of course, the class of moral patient it is not exclusively human. Indeed, it is widely understood to include non-human animals. See, for example, Peter Singer, *Animal Liberation: A New Ethics for Our Treatment of Animals* (New York: Thorsons, 1975); Tom Regan, *The Case for Animal Rights* (Berkeley: University of California Press, 1983); Mark Rowlands, *Can Animals Be Moral?* (Oxford: Oxford University Press, 2012).

[31] For opposing views, see French, *Collective and Corporate Responsibility*, p. 32; Keith Graham, 'The moral significance of collective entities', *Inquiry*, 44:1 (2001), pp. 21–41.

[32] Some theorists have maintained that the categories of moral agent and moral patient are necessarily coextensive: only moral agents can be moral patients, and all moral agents are necessarily moral patients (e.g. Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans. Mary J. Gregor and Jens Timmermann [Cambridge: Cambridge University Press, 2012]). Others argue that it is possible to be a moral patient without also being a moral agent (non-human animals being common examples) but *not* vice versa (e.g. Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, ed. J. H. Burns

Yet, importantly, the individual human constituents of formal organisations *are* moral patients and bearers of non-derivative moral rights. This, by extension, affects how we should treat institutional agents – and coheres with well-established expectations of restraint in war, for example. When a state is guilty of engaging in a war of aggression, and thereby becomes the legitimate object of defensive organised violence, there is nevertheless an expectation that fighting be conducted in a way that minimises the suffering of its population. However, the state does not have a moral right to exist for its own sake.[33]

The prospect of synthetic moral agents in world politics

Like individual human beings and formal organisations, intelligent artefacts can have sophisticated capacities for accessing and processing information, making decisions, and acting on the basis of those decisions. They lack human attributes such as consciousness and the ability to experience emotions. Yet, as compelling accounts of formal organisations as moral agents suggest, these human attributes need not be considered requisite features of moral agency. It does not follow that entities are moral agents only to the extent that they are like human beings. Given the case for including formal organisations within the class of moral agent, the possibility of artificially intelligent entities also qualifying is conceivable. This prospect would have profound implications for how we understand and explain purposive action, distribute duties, and apportion blame in world politics.

Meeting the threshold capacities for moral agency: The challenge of reflexive autonomy

If one accepts the requisite features of moral agency proposed in the first section – capacities that would allow one to understand and reflect upon moral requirements and act in such a way as to conform to them – then whether an artificially intelligent entity qualifies would seem an empirical question. Yet there is some conceptual work required to determine what meeting these criteria means in the case of intelligent machines. What hurdles must an artificially intelligent entity clear to qualify as a moral agent? In determining whether some collectivities can be moral agents, we look for the degree to which they can be considered unitary actors in their own right. An identity that does not rely on a determinate membership is important, as is an overarching decision-making apparatus that can be said to represent the will of the collectivity as a whole in a way not reducible to its constitutive parts.[34]

In the case of forms of AI, the main hurdle is autonomy, or the capacity for the entity to deliberate and act on its own. Here, though, our use of language can be misleading. Military robots, for example, are often described as acting 'autonomously' when they are programmed to identify and fire on targets without human intervention. Consider the Phalanx, for example, a navy missile system described as 'capable of *autonomously* performing its own search, detect, evaluation, track, engage, and kill assessment functions' by using a computerised radar system.[35] This is autonomy in a very weak sense. While the Phalanx is a sophisticated piece of technology, the autonomy required for an intelligent machine to qualify as a moral agent would be of a very different kind.

Intelligent artefacts could only reasonably be understood to be bearers of moral responsibilities in their own right – and be blamed for failing to discharge them – if they were able to evaluate and revise the complex codes and algorithms with which they were programmed. In other words,

and H. L. A. Hart, rev. F. Rosen [Oxford: Clarendon Press, 1996], pp. 282–3). The account of institutional moral agents defended here thereby diverges from both positions.

[33] On this point, one might consider the ethical implications of claims to state consciousness. Alex Wendt and Adam Lerner, for example, effectively engage with the possibility (and assume the appeal) of the state as a moral patient (see note 8 above), although without using this term or considering the consequences of this ethical standing.

[34] See, for example, French, *Collective and Corporate Responsibility*, p. 13; and Erskine, 'Assigning responsibilities', p. 71.

[35] United States Navy, 'MK 15: Phalanx Close-In Weapon System (CIWS)', available at: {https://www.navy.mil/resources/fact-files/display-factfiles/article/2167831/mk-15-phalanx-close-in-weapon-system-ciws/}, emphasis added.

the machines' acts and omissions could only be considered theirs if they could genuinely choose to do otherwise – and act on the basis of reasons that were their own. The plans and preferences that guide artificially intelligent entities would remain those of their individual human engineers, programmers, and operators – unless and until the machine could reflect on and alter them independently. In short, the autonomy required of intelligent machines to qualify as moral agents must be *reflexive* rather than merely reactive.[36]

Importantly, autonomy here is distinct from the freedom with respect to external conditions (noted in the first section) that a moral agent must enjoy in order to be a legitimate object of particular moral responsibility judgements. The notion of freedom alluded to above – the ability to act either *without* external constraint or coercion or *with* the necessary resources and power – is fundamental to assessing the behaviour of any moral agent. For example, to turn briefly to the moral responsibilities of restraint that will be the focus of the final section, we might excuse a soldier for failing to exercise restraint if she were ordered to shoot a civilian with a gun to her own head. Moreover, a state lacking both the material resources and independence from other agents and structures necessary to exert normative influence within its region may not be expected to discharge a responsibility to persuade a neighbouring state to refrain from attacking vulnerable populations within its borders. By autonomy, I mean something else. Autonomy entails a capacity 'to act, reflect, and choose on the basis of factors that are [the agent's] own (authentic in some sense)'.[37] In the case of the intelligent artefact, the required autonomy is an *internal* cognitive capacity that would have to be present for it to qualify as a moral agent in the first place.

Whether robots and algorithmic systems could *ever* possess this capacity for reflexive autonomy is a point on which respected scholars – from computer scientists to philosophers – would disagree. Defending a position either way is beyond the aims of this article. My goal is simply to propose criteria that such systems would have to meet to qualify as moral agents. I will refer to the as-yet-hypothetical intelligent artefacts to which one could coherently assign moral responsibilities and apportion blame as '*synthetic* moral agents'.[38] This speculative category constitutes another genuine, non-human variation on moral agency.

Unnatural and narrowly artificial agents

As for their appropriate description in relation to the 'natural' vs 'artificial' agency distinction, synthetic moral agents would be more aptly described as 'artificial' – in the sense of man-made, non-biological – than the corporate variation so often associated with this adjective. Again, Weber's vivid depiction of the latter as '*living machines*' is instructive. In contrast, whatever else intelligent machines are – and could become – they are not living. Were machines to qualify as moral agents by acquiring their own sophisticated, integrated capacities for decision making, reflexivity, and action, they would relinquish any aspect of their agency that might be described as 'natural' – outside an echo of what they may have been designed to imitate. They would be entirely artificial in a way that flesh-and-blood and institutional moral agents cannot be.

On the other hand, the extent to which they could be artificial in the specific, socially constituted sense highlighted by O'Neill is severely limited. Notably, in a very brief aside to her original

[36] I am grateful to Toby Walsh for help in formulating this distinction.

[37] John Christman, 'Autonomy in moral and political philosophy', in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy* (Spring 2020 edition), available at: {https://plato.stanford.edu/entries/autonomy-moral/}.

[38] Elsewhere, the label 'artificial moral agent' is frequently used. See, for example, Colin Allen, Gary Varner, and Jason Zinser, 'Prolegomena to any future artificial moral agent', *Journal of Experimental & Theoretical Artificial Intelligence*, 12:3 (2000), pp. 251–61; Colin Allen, Iva Smit, and Wendell Wallach, 'Artificial morality: Top-down, bottom-up, and hybrid approaches', *Ethics and Information Technology*, 7:3 (2005), pp. 149–55. Yet this label obscures the sense in which flesh-and-blood and institutional bodies are also artificial moral agents – and may inadvertently convey the notion that they are somehow less genuine moral agents. Joanna J. Bryson, Mihailis E. Diamantis, and Thomas D. Grant use the label 'synthetic persons' to address the possibility of their status as *legal* persons in 'Of, for and by the people: The legal lacuna of synthetic persons', *Artificial Intelligence and Law*, 25:3 (2017), pp. 273–91. I am inclined to avoid the label 'person' here in addressing the moral status of intelligent machines as I understand it to connote both moral agency and moral patiency.

comparison between individual and institutional agents, O'Neill turns to what she calls 'artificial intelligences' and observes that 'their capacities for cognition and action are not and cannot be extended by the complex, conceptual, material and institutional resources which structure normal human agency'.[39] This observation gestures towards a significant distinguishing feature of would-be synthetic moral agents. Algorithms represent a particular, static snapshot of the shared understandings and social mores of the individuals and institutions that create them. Intelligent artefacts thereby embody a delimited, curated, and inherently conservative version of our social world. Were synthetic moral agents possible, they would, by definition, be able to reflect on, reject, and revise these algorithmic codes. Yet they would arguably remain removed from the nuanced and dynamic — and particularly human — norms, social structures, and systems of meaning that variously enable, limit, and define the roles of flesh-and-blood and institutional moral agents.

Moral agents but not moral patients

Like institutional moral agents, synthetic moral agents would not be moral patients.[40] Acknowledging that 'robots' may be 'rational creatures', Vinit Haksar states that 'we don't feel that we owe them moral consideration for their own sakes'. Rather, 'we feel we can dismantle them and use their parts to construct more useful robots, without violating their rights; and we often adopt a similar attitude towards corporations'.[41] This illustrates well the view that intelligent machines, like formal organisations, could have instrumental value, but would lack intrinsic value.[42]

Despite this convergence, whereby neither institutional moral agents nor synthetic moral agents (if, indeed, the latter were possible) qualify as moral patients, there is another consequential difference between them. Namely, only the former are partially constituted by moral patients (in the form of their individual human members). This reality imposes limits on how we treat formal organisations despite their lack of moral standing. Attempts to punish them – for example, by imposing sanctions on delinquent states – risk directly harming their flesh-and-blood constituents instead and can thereby undermine the moral legitimacy of such actions. These limits disappear in the case of synthetic moral agents. Yet, on this point, a couple of qualification are in order. To begin, there can still arise the lesser problem of 'overspill', which has been identified in attempts to punish institutional moral agents and is understood as the *indirect* harm that results when these agents are dismantled or incapacitated and are thereby no longer able to perform certain functions upon which particular moral patients rely.[43] Overspill is also conceivable if one destroys an intelligent artefact that somehow benefits or provides a significant service to a moral patient – a robot that cares for an elderly person, for example. Moreover, it is important to note that, even if synthetic moral agents would not feel pain or suffer, the flesh-and-blood moral agents acting alongside them might believe that they do. Indeed, we already tend to make such assumptions about intelligent machines. Fascinating cases of soldiers feeling empathy towards 'injured' military robots, and demanding that they be treated with moral consideration, deserve attention.

[39] O'Neill, 'Who can endeavour peace?', p. 56.

[40] On this point, some strongly disagree. See, for example, Erica L. Neely, 'Machines and the moral community', *Philosophy & Technology*, 27:1 (2014), pp. 97–111, and John Danaher, 'Welcoming robots into the moral circle: A defence of ethical behaviourism', *Science and Engineering Ethics*, 26:4 (2019), pp. 2023–49.

[41] Vinit Haksar, *Indivisible Selves and Moral Practice* (Edinburgh: Edinburgh University Press, 1991), p. 55. Note that in this passage Haksar is denying that robots qualify as 'moral subjects' rather than using the language of 'moral patient'.

[42] It is important to note that Haksar thereby sees both robots and corporations as failing to meet the criteria for moral agency, which is at odds with my proposal here that institutional agents and (perhaps) some future forms of AI-enabled entities can be moral agents but not moral patients.

[43] Toni Erskine, 'Kicking bodies and damning souls: The danger of harming "innocent" individuals while punishing "delinquent" states', *Ethics & International Affairs*, 24:3 (2010), pp. 261–85 (pp. 274, 279). In proposing this category of collateral harm when an institutional agent is adversely affected, I was inspired by John C. Coffee, '"No soul to damn: no body to kick": An unscandalized inquiry into the problem of corporate punishment', *Michigan Law Review*, 79:3 (1981), pp. 386–459 (pp. 401–2).

One such documented case involved a robot the length of small adult, 'modelled on a stick-insect' with a multitude of legs, which was designed to tread on, and destroy, landmines, thereby protecting soldiers from risk.[44] For the engineer who created it, its test run – executed in front of a military audience – was a great success. 'Every time it found a mine, blew it up and lost a limb, it picked itself up and readjusted to move forward on its remaining legs, continuing to clear a path through the minefield.' Eventually, it had only a single remaining leg. 'Still it pulled itself forward.' At this point, however, the colonel overseeing the exercise could take no more and ordered that it be halted. 'The test, he charged, was *inhumane*.'[45]

This reaction points to a tendency to view (at least some) intelligent machines as moral patients, despite their lacking the requisite properties. One might suggest that this is all to the good. After all, even if these artificially intelligent entities would not require protection or concern for their own sake, it is conceivable that we could harm *ourselves* if we were to decline to treat them as moral patients when they present the illusion of possessing the properties that would allow them to qualify.[46] (Neither self-inflicted moral injury nor the cultivation of a habit of harming and humiliating others with disregard requires actual harm. The perception is enough.) Yet this inclination to wildly misattribute characteristics to intelligent machines should also give rise to grave concerns. If it is so easy to misperceive military robots as sentient beings (after all, the stick-insect mine-sweeper was not a particularly sophisticated example of artificial intelligence), might we not be similarly moved to misattribute other capacities – and moral *agency* – to AI-enabled entities that do not actually qualify? This misattribution of capacities to intelligent machines – including those employed in war – leads to the first of two routes to 'misplaced responsibility' that I will address in the final section, below.

Increasingly sophisticated machines: Warnings, laments, and lazy anticipations

There are currently no examples of artificially intelligent entities that meet the reflexive autonomy criterion for synthetic moral agency. Yet intelligent artefacts have generated a great deal of attention – and concern – precisely because their acquiring the degree of autonomy necessary to take them beyond human determination and control is seen by some to be a credible risk. This prospect is invoked in prominent articulations of an AI-embodied existential threat, such as Stephen Hawking's stark 2014 warning that 'the development of a full artificial intelligence could spell the end of the human race'.[47] With the apparent acceleration in AI evolution, such forebodings have become more widespread.[48] The increasing capacities of AI-enabled entities to operate independently of individual human and institutional agents also contribute to apprehensions about emerging weapons systems, with UN Secretary-General António Guterres decrying

[44] Joel Garreau, 'Bots on the ground: In the field of battle (or even above it), robots are a soldier's best friend', *The Washington Post* (6 May 2007), available at: {http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html}. I am grateful to Johanna Seibt for drawing my attention to this example. For further examples of soldiers feeling empathy towards damaged military robots, see Peter W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century* (London: Penguin, 2009), pp. 337–40.

[45] Garreau, 'Bots on the ground', emphasis added.

[46] This is reminiscent of an argument of Immanuel Kant with respect to non-human animals. See Kant, 'Of duties to animals and spirits', in Peter Heath and J. B. Schneewind (eds), *Lectures on Ethics* (Cambridge: Cambridge University Press, 1997 [1784]), pp. 212–13 (p. 212). Kant viewed non-human animals as without moral standing, existing 'only as means, and not for their own sakes, in that they have no self-consciousness' and are 'incapable of judgement'. Yet he argued that if one harms them, one 'damages the kindly and humane qualities in himself, which he ought to exercise in virtue of his duties to mankind'.

[47] Rory Cellan-Jones, 'Stephen Hawking warns artificial intelligence could end mankind', BBC News (2 December 2014), available at: {https://www.bbc.com/news/technology-30290540}; see also Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014), pp. 115–26.

[48] See, for example, Center for AI Safety, 'Statement on AI risk', 30 May 2023, available at: {https://www.safe.ai/statement-on-ai-risk}; Kevin Roose, 'A.I. poses "risk of extinction," industry leaders warn', *The New York Times* (30 May 2023), available at: {https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html}; Zoe Kleinman, 'AI "godfather" Yoshua Bengio feels "lost" over life's work', BBC News (31 May 2023), available at: {https://www.bbc.com/news/technology-65760449}.

'the weaponization of artificial intelligence' and the 'prospect of machines with *the discretion* and power to take human life'.[49] Such machines, Secretary-General Guterres asserts, are 'politically unacceptable, morally repugnant and should be prohibited by international law'.[50]

Responses to the perceived danger of increasingly automated weapons are revealing when it comes to exploring how we understand the potential capacities and moral status of future AI-enabled entities. Such responses can be organised into three categories: demands for the maintenance of what has been called 'meaningful human control', or manufactured brakes on the ability of automated systems to select and engage targets completely independently;[51] accompanying calls for a pre-emptive ban on the development, production, and use of automated weapons that could bypass such control so that this particular Pandora's box remains closed;[52] and, finally, and separately, aspirations to design lethal intelligent machines to be both 'autonomous' and able to discharge moral responsibilities.[53] The third proposal is particularly striking in light of what a robust notion of synthetic moral agency must entail.

Although the first two proposed responses to the perceived threat of increasingly automated weapons face significant practical obstacles, the third borders on conceptual incoherence. The first proposal must confront the problem that limitations placed on automated weapons in order to maintain human control will be incompatible with the increasingly high-speed, high-precision interactions that we demand of them;[54] the second optimistically relies on the unlikely prospect of all parties respecting a moratorium in what has aptly been called the latest global 'arms race'.[55] Proponents of the third response, however, are either assuming that intelligent military machines will remain securely bound by the algorithms set by their programmers, or that they could be completely autonomous yet taken for granted to adhere to what we understand to be the ethics of war. The former assumption sidesteps concerns of genuine autonomy altogether. The latter both

[49] António Guterres, 'Address to the General Assembly', 25 September 2018, available at: {https://www.un.org/sg/en/content/sg/speeches/2018-09-25/address-73rd-general-assembly}, emphasis added.

[50] António Guterres, 'Secretary-General's message to meeting of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems', United Nations Secretary General, 25 March 2019, available at: {https://www.un.org/sg/en/content/sg/statement/2019-03-25/secretary-generals-message-meeting-of-the-group-of-governmental-experts-emerging-technologies-the-area-of-lethal-autonomous-weapons-systems}.

[51] Human Rights Watch, 'Killer robots and the concept of meaningful human control: Memorandum to convention on conventional weapons (CCW) delegates', Human Rights Watch, available at: {https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control}; Heather Roff and Richard Moyes, 'Meaningful human control, artificial intelligence and autonomous weapons', Briefing Paper Prepared for the Informal Meeting of Experts on Lethal Autonomous Weapons Systems (11–15 April 2016); Richard Moyes, 'Meaningful human control', in R. Geiss (ed.), *Lethal Autonomous Weapons Systems: Technology, Definitions, Ethics, Law & Security* (Frankfurt: Federal Foreign Office, 2017), pp. 239–50.

[52] Noel Sharkey, 'The evitability of autonomous robot warfare', *International Review of the Red Cross*, 94:886 (2012), pp. 787–99; Stuart Russell, Max Tegmark, and Toby Walsh, 'Autonomous weapons: An open letter,' 28 July 2015, available at: {https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/}; Bonnie Docherty, *Making the Case: The Dangers of Killer Robots and the Need for a Preemptive Ban* (Cambridge, MA: Human Rights Watch; IHRC, 2016).

[53] Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (London: CRC Press, 2009) and 'The case for ethical autonomy in unmanned systems', *Journal of Military Ethics*, 9:4 (2010), pp. 332–41 (p. 339). Arkin not only states that his goal is to design 'autonomous unmanned systems' that would 'comply with the restrictions of international law' and 'the ideals enshrined within the Just War tradition' but also suggests that confronting the challenge of 'ensuring moral performance' would involve 'reflective … processing', thereby gesturing towards the criterion for genuine autonomy discussed above. See Arkin, 'The case for ethical autonomy', p. 339.

[54] Indeed, there is evidence that some weapons systems have already exercised independence from human control. A recent UN report claims that an airstrike against Libyan National Army forces, conducted from the spring of 2020 by Libya's Government of National Accord, was the result of lethal autonomous weapons systems – STM Kargu-2 drones – operating in fully autonomous mode. See United Nations, 'Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011)', para. 63, available at: {https://undocs.org/S/2021/229}; and Stuart Russell, Anthony Aguirre, Emilia Javorsky, and Max Tegmark, 'Lethal autonomous weapons exist; they must be banned', IEEE Spectrum (16 June 2021), available at: {https://spectrum.ieee.org/lethal-autonomous-weapons-exist-they-must-be-banned}.

[55] Tom Simonite, 'For superpowers, artificial intelligence fuels new global arms race', *Wired* (8 September 2017), available at: {https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/}. See also Russell, Tegmark, and Walsh, 'Autonomous weapons'.

overlooks that the reflexive autonomy required for moral agency entails the possibility of *choosing* to violate moral norms, and also neglects one of the fundamental points that this article aims to highlight: that the unique constellation of characteristics that define different categories of moral agency must have consequences for how we can reasonably expect agents within each to act. Our failure to anticipate a distinct category of moral agent – one to which the differences between flesh-and-blood and institutional moral agents should alert us – leads to the second potential variation on 'misplaced responsibility' to be addressed in the final section, to which I will now turn.

Moral agents of restraint and the problem of misplaced responsibility in war

AI is infiltrating every domain of world politics, yet its impact on war has the potential to be particularly consequential. I will continue to invoke the backdrop of war in what follows to illustrate the practical implications of failing to define and distinguish between different categories of moral agent when describing, prescribing, and evaluating acts and outcomes involving intelligent artefacts. Specifically, I will identify two instances of misplaced responsibility. Each arises when our expectations regarding particular duties become unmoored from an accurate identification of moral agents able to discharge them.

Fundamental to the practice of war is an iterated and influential discourse on moral responsibility. Insofar as we accept that organised violence remains within the realm of morality, the acts and omissions of those that participate in it are judged against norms that dictate when it is permissible and prohibited to engage in war, and what is just and unjust in its conduct. Permission to take the first, collective step into battle awaits a justifying cause: self-defence of the state, or, perhaps, the protection of vulnerable peoples beyond its borders. Choices of weapons and targets are constrained by principles established to minimise suffering and deaths – principles that demand identifying and categorising persons and properties and carefully weighing necessity, likely harm, risk, and the relative value of individual human lives. In short, bound up in the ethics of war is a powerful expectation that actors exercise restraint.

The most influential framework for such appeals to duties of forbearance, and charges of blame when these are abrogated, is the just war tradition: an evolving consensus on principles to guide appropriate behaviour in the context of organised violence.[56] These principles have been considered, contested, and refined over centuries, are codified in international law, and outline what is understood to be morally permissible and prohibited in war. Although they allow both engagement in war and violent conduct within it to be justified if particular conditions are met, they also place a heavy burden of restraint on the participants (and potential participants) in armed conflicts. These established norms of restraint are conventionally organised into two categories: *jus ad bellum* principles, which both license and limit the resort to organised violence, and *jus in bello* principles, which serve simultaneously to condone and curtail conduct within it. Prominent principles within each include, respectively, the responsibility to refrain from the resort to organised violence in the absence of a 'just cause',[57] and the duty to discriminate between combatants and non-combatants once the fighting has begun.[58] Such principles of restraint prescribe what actors

[56] For a concise overview of contemporary just war thinking, see Seth Lazar, 'War', in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020 edition), available at: {https://plato.stanford.edu/entries/war/}. An essential reference point for contemporary just war thinking remains Michael Walzer, *Just and Unjust Wars: A Moral Argument with Historical Illustrations*, 3rd ed. (New York: Basic Books, 1992).

[57] 'Just cause' is narrowly defined as defence of the state against aggression and increasingly extended to include the protection of vulnerable populations from mass atrocity crimes when their own state manifestly fails to protect them or, indeed, constitutes the threat. By this extension I am referring to action taken in accordance with the 'responsibility to protect', which was endorsed by all member states of the United Nations at the 2005 World Summit (see United Nations, 'World Summit outcome: General Assembly resolution 60/1', 24 October 2005, available at: {https://undocs.org/en/A/Res/60/1}, and Ban Ki-Moon, 'Implementing the responsibility to protect', 12 January 2009, available at: {https://undocs.org/en/A/63/677}).

[58] According to this principle, combatants may be targeted, but harm to non-combatants is only permissible as an unintended and proportionate side effect of justified attacks on military targets.

should do and refrain from doing both in the resort to armed conflict and in its conduct. Relatedly, they are invoked to evaluate agents' acts and omissions after armies have been deployed or individual shots have been fired. In short, they are widely endorsed standards for judging both prospective and retrospective moral responsibility in war.

Identifying moral agents of restraint

In war, as in other contexts, the assignment of such moral responsibilities – and the apportioning of blame when they are abrogated – must be directed towards moral agents to be meaningful. I will use the label 'moral agent of restraint' for a body that possesses the sophisticated, integrated capacities for deliberation, reflexivity, and action necessary to qualify as a moral agent and which has some role or influence in the decisions and actions related to either the resort to organised violence or its conduct.[59] When it comes to those moral agents of restraint with which we are currently familiar, we reasonably expect them to discharge at least some of the duties to exercise or promote forbearance encompassed within the just war tradition – namely those commensurate with the agents' capacities and roles – given enabling conditions.

There are numerous examples of flesh-and-blood and institutional moral agents of restraint. With respect to the former, soldiers are the most obvious moral agents of restraint. Morally responsible for the acts and omissions within their control, they are traditionally tasked with upholding the *jus in bello* principles of 'non-combatant immunity' and 'proportionality'. Commanding officers are moral agents of restraint whose particular roles are deemed to be accompanied by additional responsibilities for ensuring that those acting under them understand and abide by the rules of war. These additional responsibilities are derived from their power to influence the cultures and practices of the collectivities they lead. Most adult citizens within liberal democracies are also moral agents of restraint. According to Michael Walzer and others, such citizens bear responsibilities not only to vote for governments that they are confident will not prosecute unjust wars, but also to speak out, to march, to protest, and to hold their governments to account if they do.[60] Moreover, political leaders such as Joe Biden, Vladimir Putin, and Xi Jinping are flesh-and-blood moral agents of restraint. They each have the influence, resources, and access to comprehensive (often privileged) intelligence – not to mention immensely destructive weapons – that come with the role of head of a powerful state and bolster both their individual human capacities for deliberation and action and what we can reasonably expect of them. In other words, they are artificial flesh-and-blood moral agents of restraint *par excellence*.

Institutional moral agents of restraint include, most prominently, the majority of states, which possess powerful capacities for waging war and accompanying *jus ad bellum* responsibilities to do so only under certain conditions. IGOs such as the United Nations (UN), the North Atlantic Treaty Organization (NATO), and the African Union are also potentially moral agents of restraint. They are expected to limit their resort to force to cases of collective self-defence or to instances of human protection in which all pacific options have been exhausted. By contrast, neither so-called failed states nor informal associations of states such as G7 and G20 groupings and 'coalitions of the willing' are moral agents of restraint. They lack the formal organisational structure and decision-making procedures that would allow them to deliberate and act at the corporate level and thereby qualify as moral agents.[61] This means, simply, that the locus of responsibility lies elsewhere. Acts and forbearances in the exercise of organised violence are more accurately described, prescribed, and evaluated at the level of the individual and institutional agents that constitute them. The states

[59] This label is inspired by O'Neill's notion of 'agents of justice' (Onora O'Neill, 'Agents of justice', *Metaphilosophy*, 32:1–2 [2001], pp. 180–95).

[60] Walzer, *Just and Unjust Wars*, pp. 296–303; see also David Estlund, 'On following orders in an unjust war', *Journal of Political Philosophy*, 15:2 (2007), pp. 213–34 (p. 234); Neta C. Crawford, 'War "in our name" and the responsibility to protest: Ordinary citizens, civil society, and prospective moral responsibility', *Midwest Studies in Philosophy*, 38:1 (2014), pp. 138–70.

[61] For discussions of why so-called failed states and informal associations of states do not qualify as moral agents, see Erskine, 'Assigning responsibilities', p. 79 and 'Coalitions of the willing', pp. 120–5.

and formal groups of states that do qualify as moral agents of restraint arguably also bear a responsibility to petition and persuade other institutional agents to refrain from engaging in wars of aggression. This is analogous to the responsibility to protest that Walzer and others maintain is borne by the individual citizen in a liberal democracy – and can arguably be discharged to greater effect.

What about synthetic moral agents of restraint? Intelligent machines increasingly contribute to war. Yet, as has already been addressed, they cannot currently be considered moral agents. There are no synthetic moral agents of restraint – only AI-enabled tools employed by flesh-and-blood and institutional moral agents of restraint. Worryingly, however, the latter are inclined to behave as if this were not so. The conception of misplaced responsibility in war that I will introduce below results from our faulty expectations of the entities that we (sometimes mistakenly) identify as moral agents of restraint. This potentially catastrophic misalignment between our responsibility judgements and the objects of such assessments can take at least two forms. Both threaten to accompany the arrival – or *perceived* arrival – of synthetic moral agents of restraint on the battlefield and in the war room. The first is already apparent and immediately consequential. It arises when we try to assign moral responsibilities to those entities that do not qualify as moral agents at all. The second is speculative. It involves our failure to distinguish adequately between entities that occupy different categories of moral agent and our corresponding inability to calibrate our expectations of them according to their markedly different capacities and limitations. I will address each in turn.

Immediate concerns: Abdicating responsibilities to non-moral agents

Moral responsibility is often actively eschewed. We have a tendency to disown responsibility both in the prospective sense by refusing to accept that a particular burden to act (or refrain from acting) is ours, and in the retrospective sense by denying that we are the ones who are blameworthy. Saying – both to ourselves and others – that another agent is answerable in our stead eases what is aptly described as the 'weight' of responsibility. With this weight ostensibly borne elsewhere, one can take a deep breath, disengage, look away, and maintain a clear conscience. One particularly problematic way of achieving this desired unburdening is to redirect responsibilities and deflect blame to an entity that is not a moral agent at all. A common instance of this occurs in international politics when duty or blame is asserted obliquely to lie with the 'international community', an amorphous collectivity incapable of unified, purposive action, rather than directed towards, or assumed by, relevant institutional moral agents that are able to answer specific calls to action and charges of wrongdoing.[62]

There is also a tendency to assume that the weight of responsibility is somehow reduced when it is shared with other agents. Michael Barnett, for example, laments our propensity for 'democratizing blame' in international politics, citing cases of guilty parties identifying a multitude of other agents that purportedly share responsibility for some wrongdoing in an attempt to reduce 'their own particular culpability to a meaningless fraction'.[63] While an agent's moral responsibility (in either the prospective or retrospective sense) need not be understood as diminished when it is shared – there are more demanding ways of understanding 'shared responsibility'[64] – worrying scenarios arise when responsibility is seen to be reduced because it is purportedly shared with an entity that cannot be expected to bear a burden of moral responsibility at all. In such cases, responsibility is understood to be divided and distributed, yet there is no other moral agent capable of bearing the apportioned weight. The result is the sole duty-bearer's perilous misperception that her burden is lessened because the task of 'doing the right thing' is being jointly undertaken and

[62] Erskine, 'Assigning responsibilities', p. 73 and 'Coalitions of the willing', pp. 117–18, 120.

[63] Michael Barnett, *Eyewitness to a Genocide: The United Nations and Rwanda* (Ithaca, NY: Cornell University Press, 2002), p. 154.

[64] See, for example, Erskine, 'Coalitions of the willing', pp. 134–5. I return to this point in 'Mitigating the risks of misplaced responsibility in war', below.

any blame for harm or wrongdoing will not be hers alone. Tempting circumstances for assuming that the moral responsibility to exercise restraint is either redirected or diminished arise when flesh-and-blood and institutional moral agents of restraint rely on AI-enabled automated weapons and decision-support systems in war.

Misperceptions of machine moral agency

The speed of machine cognition is rapidly surpassing that of human beings in certain domains. Moreover, machine-learning processes are necessarily opaque and often unpredictable. Those who operate and are guided by intelligent machines often do not understand how these machines make decisions and frequently fail to grasp their limitations. The lack of transparency in AI-driven decision making can have a range of negative consequences across many different contexts. Algorithms can covertly perpetuate inequality and reinforce society's biases, with implications for recruitment and insurance decisions, policing practices, and the allocation of welfare.[65] Furthermore, algorithmic risk assessments that predict recidivism, for example, and guide the decisions of judges during sentencing are open to charges that they violate due process because consequential judgements can neither be explained nor defended.[66] These are grave problems. Yet this algorithmic opacity – combined with particular human tendencies and biases – also risks affecting both how the capacities of intelligent machines are perceived (or misperceived) and the self-perception of the responsible agents that use these tools. These agents may feel removed from crucial decisions and actions – and less answerable for their consequences. This effect could have particularly grave repercussions in the context of moral responsibilities of restraint in war.

Human actors are affected by 'automation bias', or a tendency 'to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct'.[67] Discussing the implications of introducing automation to decision-support systems embedded in computer interfaces, Mary L. Cummings identifies the risk of a perceived reduction in the human user's own 'sense of moral agency and responsibility'.[68] We tend to see an automated system 'as an independent agent capable of wilful action'.[69] Cummings warns of the resulting creation of a 'moral buffer', which 'allows people to ethically distance themselves from their own action'.[70] We erroneously see ourselves displaced by machines as the relevant moral agents. Problematically, this tendency can encourage us to interpret our military tools as sites of legitimate authority and loci of responsibility.[71]

Relinquishing responsibility for restraint

The use of AI-enabled automated weapons ushers automation bias onto the battlefield. Consider South Korea's Super aEgis II. This robotic sentry, in the form of an automated turret, was designed for and tested in the demilitarised zone between North and South Korea and has been exported for

[65] See, for example, Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (New York: St. Martin's Press, 2018); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown Publishers, 2016).

[66] Frank Pasquale, 'Secret algorithms that threaten the rule of law', *MIT Technology Review* (1 June 2017).

[67] Mary L. Cummings, 'Automation and accountability in decision support system interface design', *Journal of Technology Studies*, 32:1 (2006), pp. 23–31 (p. 25). See also Linda J. Skitka, Kathleen L. Mosier, and Mark Burdick, 'Does automation bias decision-making?', *International Journal of Human–Computer Studies*, 51:5 (1999), pp. 991–1006; and Kathleen L. Mosier and Dietrich Manzey, 'Humans and automated decision aids: A match made in heaven?', in Kathleen L. Mosier and Dietrich Manzey (eds), *Human Performance in Automated and Autonomous Systems* (Boca Raton, FL: CRC Press, 2019), pp. 19–42.

[68] Cummings, 'Automation and accountability', p. 23.

[69] Cummings, 'Automation and accountability', p. 28.

[70] Cummings, 'Automation and accountability', p. 23; see also Batya Friedman and Peter H. Kahn Jr, 'Human agency and responsible computing: Implications for computer system design', *Journal of Systems Software*, 17:1 (1992), pp. 7–14.

[71] See, for example, Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W.W. Norton and Company, 2018), pp. 276–9; Elke Schwarz, *Death Machines: The Ethics of Violent Technologies* (Manchester: Manchester University Press, 2018), pp. 158–9.

use in other countries, including the United Arab Emirates and Qatar.[72] Although it has the capacity to identify, track, and shoot targets entirely independently of human mediation, in its current incarnation the Super aEgis II is built to require a human operator to provide authorisation before a shot is fired.[73] After the Super aEgis II identifies a target, the human operator provides this authorisation by first entering a password in a nearby computer to 'unlock the turret's firing ability' and then providing the manual input required for it to shoot.[74] Yet, with such weapons systems, ostensible safeguards of 'human in the loop' and 'human on the loop' mechanisms (which, respectively, require human authorisation to fire and entail human oversight and override provisions) are undermined if human operators tend to accept uncritically the machine's automated selection of targets. In other words, the flesh-and-blood moral agent of restraint effectively removes herself from the loop in terms of accepting responsibility for what remain her decisions and actions. Algorithms that rely on big data analytics and machine learning to recommend targets (by uncovering correlations in large amounts of data drawn from individuals' text messages, web browsing, email, and location), such as those used by the United States for drone strikes in Yemen and Pakistan, and by Israel for bombing in Gaza,[75] similarly threaten to affect how the human agents who rely on them perceive their own roles. In both cases, the human operator risks seeing herself as the obedient recipient of instructions rather than the ultimate decision maker. *Jus in bello* responsibilities are thereby deflected.

Separately, the use of machine-learning algorithms to advise governments on resort-to-force decisions is unlikely to be far away.[76] Intelligent decision-support systems that could guide the initiation of hostilities – by, for example, estimating threats, anticipating potential adversaries' movements, forecasting casualties, and predicting mission costs – would risk introducing automation bias into the war room and *jus ad bellum* considerations. There is no reason to think that institutional decision making would not be similarly susceptible to automation bias.[77] The corporate decision-making bodies of states and IGOs could likewise see themselves as relinquishing responsibility for restraint as they defer to algorithms in the decision to go to war.

[72]Simon Parkin, 'Killer robots: The soldiers that never sleep', BBC Future (17 July 2015), available at: {http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-that-never-sleep}; Erico Guizzo and Evan Ackerman, 'Do we want robot warriors to decide who lives or dies?', IEEE Spectrum (31 May 2016), available at: {https://spectrum.ieee.org/robotics/military-robots/do-we-want-robot-warriors-to-decide-who-lives-or-dies}; Vincent Boulanin and Maaike Verbruggen, 'Mapping the development of autonomy in weapon systems', Solna, Stockholm International Peace Research Institute, 2017, pp. 1–131 (pp. 44–7), available at: {https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf}.

[73]Parkin, 'The soldiers that never sleep'.

[74]Parkin, 'The soldiers that never sleep'.

[75]John Naughton, 'Death by drone strike, dished out by algorithm', *The Guardian* (21 February 2016), available at: {https://www.theguardian.com/commentisfree/2016/feb/21/death-from-above-nia-csa-skynet-algorithm-drones-pakistan}; Jennifer Gibson, 'Death by data: Drones, kill lists and algorithms', *E-International Relations* (18 February 2021), available at: {https://www.e-ir.info/2021/02/18/death-by-data-drones-kill-lists-and-algorithms/}; Harry Davies, Bethan McKernan, and Dan Sabbagh '"The Gospel": How Israel uses AI to select bombing targets in Gaza', *The Guardian* (1 December 2023), available at: {https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets}.

[76]Ashley Deeks, Noam Lubell, and Daragh Murray, 'Machine learning, artificial intelligence, and the use of force by states', *Journal of National Security Law and Policy*, 10:1 (2018), pp. 1–25 (p. 2). See also Toni Erskine and Steven E. Miller, 'AI and the decision to go to war: Future risks and opportunities', *Australian Journal of International Affairs*, 78:2 (2024).

[77]Most empirical studies of automation bias have focused on single-person cases. Yet a few studies have demonstrated the persistence of automation bias in teams. See, for example, Linda J. Skitka, Kathleen L. Mosier, Mark Burdick, and Bonnie Rosenblatt, 'Automation bias and errors: Are crews better than individuals?', *The International Journal of Aviation Psychology*, 10:1 (2000), pp. 85–97; Kathleen L. Mosier, Linda J. Skitka, Melisa Dunbar, and Lori McDonnell, 'Aircrews and automation bias: The advantages of teamwork?', *The International Journal of Aviation Psychology*, 11:1 (2001), pp. 1–14; and Kathleen L. Mosier and U. M. Fischer, 'Judgment and decision making by individuals and teams: Issues, models, and applications', *Reviews of Human Factors and Ergonomics*, 6:1 (2010), pp. 198–256. One would not only expect automation bias in both single-person and team scenarios to contribute to errors in organisational decision making, but it is possible that automation bias could also directly affect specifically organisational decision making. (This possibility warrants further study.)

In sum, flesh-and-blood and institutional moral agents may believe that they are off the moral hook when sophisticated military machines are mistaken for moral agents. Our tools become our moral proxies, our moral guides and compasses, and our scapegoats. Moral responsibility is thereby misplaced – and we are diminished.

Future considerations: Eliding distinct categories of moral agent

AI-driven systems that display varying degrees of qualified autonomy can only be considered the tools of the agents that employ them. Moral responsibilities to exercise restraint – and blame when these are derogated from – remain with these flesh-and-blood and institutional moral agents. Yet it is imaginable that we may, one day, create a new form of moral agent. However remote this prospect, the complications that would accompany it deserve attention – if only to highlight the folly in assuming that successfully designing synthetic moral agents would ensure that what we understand to be moral responsibilities (including responsibilities of restraint in war) would be discharged. After all, it is likely that this new variant of moral agent would depart in significant ways from our eminently imperfect human ideal. This lesson can be drawn from the comparative analysis of flesh-and-blood and institutional moral agents. This should, in turn, lead us to consider whether intelligent artefacts, even if they could qualify as moral agents at some future point, would prove a difficult fit with our own substantive moral codes. The ethics of war provides a sobering context for considering this misalignment.

If intelligent machines could ever become autonomous in the strong, reflexive sense required for moral agency – and this remains a resounding *if* – their particular embodiment would likely be very different from the flesh-and-blood and institutional embodiments with which we are most familiar. Synthetic moral agents would be silicon and software. They would be unaffected by emotions and neither empowered nor constrained by our social structures and systems of meaning. Unthreatened by the vulnerabilities of biological life, they would be impervious to pain. These differences would have consequences for the ethical reasoning to which they could reasonably be expected to respond, even if they were to achieve the threshold capacities necessary to qualify as moral agents. For example, while our capacities for deliberation, reflexivity, and action allow us to engage in ethical reasoning and discharge particular duties, certain vulnerabilities inform our conception of what these duties are. A duty to minimise suffering is compelling because we know what it is to feel pain.

Moral agency, I have argued above, need be accompanied by neither human characteristics nor the status of moral patient.[78] Yet what would the synthetic moral agent's proposed lack of emotion, immunity to social expectations, and invulnerability to pain and suffering mean for moral motivation as we understand it? Perhaps the absence of emotions would inoculate it against acrimony towards 'the enemy' and any accompanying temptation to commit atrocities. The passionless bot may be a paragon of impartiality. Yet it would also be incapable of empathy, and, for better or worse, empathy and 'fellow-feeling' with the enemy have played a significant role in just war thinking, as has the sentiment of a common humanity, however circumscribed in application. Moreover, if the synthetic moral agent were not defined by our social roles and practices, as I have suggested, it would be unconstrained by the often-powerful norms that accompany them. Without a desire to be recognised within this community, and be seen to conform to shared conceptions of appropriate conduct, espoused responsibilities of restraint would carry less force.

[78]For a very different view, see Robert Sparrow's pioneering article on questions of moral responsibility in relation to AI-enabled weapons in war, 'Killer robots', *Journal of Applied Philosophy*, 24:1 (2007), pp. 62–77. Sparrow maintains that we must be able to punish or reward an AI-enabled system for it to make sense to hold it morally responsible for its actions, which he takes to mean that it must be able 'to suffer' (pp. 71–2). In other words, he assumes that an autonomous weapons system could only qualify as a moral agent if it were also a moral patient. By contrast, I have argued that moral patiency is a defining characteristic of individual human moral agents rather than a requisite feature of moral agency itself. Furthermore, my position is that neither the coherence nor practical value of attributing moral responsibility to an agent rests on the possibility of punishing it.

Finally, invulnerability to pain and suffering would preclude an alternative motivation for restraint grounded in considerations of reciprocity. In the absence of both fellow-feeling with a human enemy and susceptibility to the social constraints of international norms, conduct in war would be unlikely to be curtailed, instead, by the synthetic moral agent's self-interested desire to be the beneficiary of restraint. If intelligent machines were to qualify as moral agents in their own right, they would (according to just war principles) owe flesh-and-blood agents duties of restraint, yet they would not be the sorts of entity to which we would owe moral consideration in return.[79] And, even if we were to exercise the type of restraint required by international humanitarian law in our interactions with military robots, thereby abiding by rules of restraint formulated to limit suffering, the gesture would lack value for synthetic moral agents and so fail to provide a rational basis for self-interested reciprocity.

My modest point here is that it is worth considering whether the type of ethical reasoning that we demand of moral agents of restraint in war requires a particularly human variation on moral agency – with all of the baggage and imperfections that this entails. Sophisticated, integrated capacities for deliberation, reflexivity, and action might (in a possible future world) allow artificially intelligent entities to qualify as moral agents, but this may not be enough to enable them to respond to, and replicate, *our* ethical reasoning. Their initiation into our moral universe – as it is currently conceived – may prove impossible.

Mitigating the risks of misplaced responsibility in war

In the words of Alan Turing, writing almost 75 years ago about conceivable 'thinking machines' of the future, 'we can only see a short distance ahead, but we can see plenty there that needs to be done'.[80] Whether or not autonomous robots, algorithmic systems, and other machines have the potential to achieve genuine moral agency at some point in the future, in their current form they often masquerade as moral agents. In response to the accompanying risk of abdicating our moral responsibilities to them, we might consider some preventative measures.

With respect to the error of seeing one's moral responsibility for particular acts and outcomes as diminished when one relies on an ostensible moral agent (that does not, in fact, qualify as such), we would do well to rethink how we attribute responsibility to actual moral agents that, respectively: (i) act in concert; and (ii) have their capacities augmented by AI-enabled tools. To begin, when multiple moral agents act in concert to contribute to an outcome that none could achieve acting independently, the responsibility borne by each is *not*, in fact, diminished. Such *shared responsibility* – or moral responsibility that is necessarily distributive amongst individual contributors to an outcome that could only result (or have resulted) from their deliberately acting together in pursuit of a common goal – does not reduce the moral burden of any.[81] Rather, each moral agent is responsible for the act or outcome to which they either can (prospectively) contribute (in terms of discharging a duty) or have (retrospectively) contributed (in terms of warranting praise or blame).[82]

[79] This is different in an important respect to institutional moral agents, which, although they are not moral patients at the corporate level, are constituted by moral patients, thereby (in theory) allowing for meaningful reciprocity in the exercise of restraint.

[80] Alan M. Turing, 'Computing machinery and intelligence', *Mind*, 59:236 (1950), pp. 433–60 (p. 460).

[81] I offer this conception of 'shared responsibility', linked to a notion of 'joint purposive action', in Erskine, 'Coalitions of the willing', pp. 134–5. (The concept of 'joint purposive action' is inspired by Larry May's notion of 'joint purposive behaviour' but departs from it in important respects. See May, *The Morality of Groups*, p. 26.) The 'individual contributors' to which I am referring here might be flesh-and-blood or institutional moral agents.

[82] Note that this conception of shared moral responsibility being distributed amongst relevant moral agents in the context of joint purposive action is very different from, for example, Luciano Floridi, 'Faultless responsibility: On the nature and allocation of moral responsibility for distributed actions', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374:2083 (2016), pp. 1–13. For Floridi, non-intentional contributions to a morally loaded outcome by a 'network of agents' (including machines) should result in 'moral responsibility' being pragmatically distributed 'fully' to each 'agent' of this network, yet deemed 'faultless' (a seeming contradiction). His account provocatively renders questions of moral

This means that *even if* one (as a moral agent of restraint) mistakes the AI-enabled system alongside which one is acting for a moral agent in its own right, the default assumption should not be that some degree of responsibility has been lifted from one's own shoulders.

Separately, even though shared moral responsibility cannot encompass non-moral agents such as AI-enabled automated weapons and decision-support systems,[83] complex forms of machine–human (and machine–institution) interaction and teaming *should* affect our judgements of moral responsibility. After all, AI-enabled tools bolster the cognitive, decision-making, and executive capacities of both flesh-and-blood and institutional moral agents. Drawing on O'Neill's work, above, I highlighted the reality of individual human capacities being extended by certain social roles and thereby achieving an 'artificial' dimension. Here is yet another way in which it is possible to augment what O'Neill describes as agents' 'capacities to act and abilities to foresee'.[84] Flesh-and-blood and institutional moral agents can acquire an additional, 'artificial' dimension with the aid of intelligent machines. Importantly, with the capacities of these moral agents thereby enhanced, our expectations of them must change. Our expectations of the soldier who is operating an AI-enabled automated weapon, for example, and her expectations of herself, should increase proportionately to her enhanced capacities to deliberate and act – as should our expectations of the state employing a decision-support system to evaluate the permissibility of the resort to force. In short, their AI-enhanced capacities *magnify* their moral responsibilities of restraint.

In addition to rethinking how we see the responsibilities of moral agents who employ and are aided by AI-enabled tools, we might also propose some *supplementary* moral responsibilities of restraint in order to mitigate the risk of abdicating our responsibilities to entities that lack moral agency. By these, I mean responsibilities to create the conditions within which our commonly propounded duties of forbearance in war can be effectively discharged. These might include responsibilities to design automated weapons and decision-support systems in such a way that they cannot easily be mistaken for moral agents in themselves and therefore do not function as 'moral buffers' (to return to Cummings's phrase). This could involve, for example: making the decision making of machine intelligence, along with its limitations, more transparent; actively discouraging the misperception of machine moral agency by refraining from anthropomorphising intelligent artefacts; and incorporating cues into such machines that reinforce human agency and responsibility.

On the last point, an example of such a cue – even if it was not intended as such – brings us back to South Korea's Super aEgis II. After identifying a potential target and before receiving the go-ahead to fire from the human operator, the Super aEgis II issues a warning by broadcasting (in Korean): 'Turn back. Turn back, or *we'll* shoot!'[85] The plural subject invoked in this command explicitly co-opts the 'human in the loop' operator into the decision making and lethal action. In other words, the statement effectively serves not only to warn any trespassers,[86] but also to remind the human operator of her direct role in the decision to use lethal force, thereby countering any

agency unnecessary. The types of activity referred to here, by contrast, involve moral agents purposefully coming together to contribute to a common goal and moral responsibility attributed to each of them individually in a way that is not qualified.

[83] Shared responsibility according to this account could not encompass intelligent machines unless and until synthetic moral agency were possible.

[84] O'Neill, 'Agents, agencies and responsibility', p. 15.

[85] Parkin, 'The soldiers that never sleep', emphasis added. According to Parkin, the Super aEgis II must always be accompanied by an 'acoustic hailing robot' (in the form of a large speaker on a tripod), which allows the Super aEgis II to broadcast this warning.

[86] This is the intended function of the broadcast: to discharge another *jus in bello* responsibility of restraint; namely to take 'due care'/'constant care' and 'all feasible precautions' (including by issuing 'effective advance warning') in order to minimise incidental civilian deaths. For an account of this moral responsibility, see Walzer, *Just and Unjust Wars*, pp. 151–6. As it is enshrined in international law, see Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, art. 57(1) and 57(2)aii and 57(2)c, available at: {https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-57}. Of course, this responsibility of restraint is thereby discharged by the individual human and institutional moral agents that design, deploy, and operate the Super aEgis II and *not* by the machine itself.

imagined 'moral buffer'. (More appropriate still – although unwieldy – would be: '*Stop or my human counterpart will decide to shoot!*') In short, such a cue signals, and reinforces, the appropriate locus of responsibility.

Finally, with respect to the tendency to misplace responsibility when we look to the future and combine confidence in our ability to create reflexively autonomous intelligent systems with a failure to consider the duties that we could reasonably expect these creations to discharge, we might heed a simple principle. Namely, any compelling attribution of moral responsibility must be informed by the specific capacities and limitations of the entity towards which it is directed. Expectations of restraint directed towards as-yet-hypothetical synthetic moral agents, but based on our understanding of human capacities and limitations, are fanciful at best and reckless at worst. Countering this instance of misplaced responsibility requires guarding against hubris – and embracing preliminary lessons from this comparative analysis of existing and potential moral agents in world politics.

Conclusion

AI-enabled entities will continue to intervene in world politics over the next 50 years – including in ways that we cannot yet imagine. Reassessing our often-implicit assumptions about sophisticated forms of agency in IR is necessary in order to develop the theoretical framework required to describe, prescribe, and evaluate the decisions and actions with which these emerging technologies are associated. As a way of beginning to address the capacities, relative autonomy, and status of AI-enabled entities, I have proposed a preliminary typology of moral agency in world politics, which compares and contrasts 'flesh-and-blood', 'institutional', and 'synthetic' variations. This analysis yields valuable insights – and points to areas for future study.

Significantly, this analysis demonstrates that there are different embodiments of purposive actor that can potentially meet the qualifying criteria for moral agency, as important scholarship across a range of theoretical perspectives in IR has intimated (although not always explicitly acknowledged). Moreover, these embodiments have distinct characteristics, capacities, and limitations. Two lessons follow. Each has important implications for thinking about the prospect of synthetic moral agency – and for disentangling different conceptions of duty bearer when we make weighty attributions of moral responsibility in world politics.

The first lesson drawn from this analysis is that there is a crucial distinction between the threshold capacities for moral agency and the characteristics that define particular categories of moral agent in world politics. Although looking to human beings may help us to distil what we understand to be the defining features of bodies to which we can assign responsibilities and apportion blame, it does not follow that human beings either exhaust the category of moral agent or represent an ideal against which other potential variations should be judged. Compelling, non-human variations on moral agency are already widely recognised in both IR theory (albeit often implicitly) and in practical discourses on international politics. Both individual human beings and formal organisations can possess the capacities necessary to qualify as moral agents. One of the implications of this observation is that it should not be deemed impossible for intelligent machines to achieve the status of moral agent simply because they lack human characteristics. Indeed, this somewhat unsettling provocation was one of the prompts for this article. However, another implication of this observation is that it is a fundamental error to assume, as we are inclined to do, that intelligent machines inch closer to being able to bear the burdens of duty and blame the more human they appear. For example, language-generative models that use predictive machine-learning techniques to string together words and create the impression of human-like empathetic engagement are not synthetic moral agents. Nor are AI-enabled decision-support systems appropriate sites for our individual or institutional unburdening of responsibility because they seem to take on authoritative decision-making roles when they predict possible threats or recommend targets for drone strikes. Moreover, even though the AI-enabled Super aEgis II robotic sentry appears laudable in issuing a spoken warning after independently identifying a potential target, and before receiving

the go-ahead to fire from its human operator, it cannot coherently be considered a moral agent of restraint. When it comes to synthetic moral agency, reflexive autonomy is a fundamental qualifying feature. As such, no autonomous robot, algorithmic system, or other intelligent artefact can currently qualify. Loci of responsibility cannot (yet) include intelligent machines.

A second lesson, which follows from the insight that even bodies that share the sophisticated decision-making, reflexive, and executive capacities that would allow them to qualify as moral agents may have radically different defining features, is that these differences must be taken into account in our attributions of moral responsibility. Even a body that qualifies as a moral agent cannot be expected to perform actions for which it does not possess the requisite capacities. We cannot reasonably have the same expectations of different categories of moral agent. This is borne out by how we distribute moral responsibilities between individual human and institutional agents in terms of their relative capacities and limitations (through the attribution of *jus in bello* responsibilities of restraint to soldiers and *jus ad bellum* responsibilities of restraint to states in traditional just war theorising, for example). Any future realisation of the category of synthetic moral agent would also have its own unique defining features. Anticipating the characteristics, capacities, and limitations of such entities based on a comparative analysis of existing moral agents counters the lazy assumption that they would simply reproduce our individual human example. It also solicits caution about what we could hold such entities answerable for.

Problems arise when these two lessons are not grasped. I have described resulting missteps as instances of 'misplaced responsibility'. Although I have introduced them here against the backdrop of organised violence, where their effects are particularly stark and potentially devastating, they are equally applicable to other global contexts where individual human and institutional agents employ and interact with AI-enabled entities. These other contexts might involve, for example: global financial transactions; the creation, curation, and distribution of information (including to influence domestic and foreign elections); the analysis and processing of refugee claims; and the prediction of climate crises with accompanying recommendations for remedial responses and the allocation resources.

The first instance of misplaced responsibility involves assuming that moral agents exist where they do not. When it comes to AI-enabled entities in world politics, this is a neglected risk that we need to confront now. It has always been tempting to disown one's choices and redirect responsibility in challenging circumstances – in war, through the (illegitimate) appeal to 'superior orders', for example.[87] Yet the cloak of algorithmic authority and the misperception that intelligent machines possess capacities that they do not, combined with our own human tendency to try to reduce the weight of the responsibility that we bear, have together created a new moral crutch in AI-enabled military tools – one less understood and ostensibly more legitimate. Moreover, when the seemingly authoritative decisions and apparently autonomous (but in fact precisely programmed) actions of these intelligent machines can neither be interpreted nor audited by those who deploy, operate, and are guided by them, they become even more likely to be blindly accepted. In such cases, an engineered, algorithmic 'fog of war' threatens to obscure appropriate sites of moral responsibility.

The second instance of misplaced responsibility entails failing to distinguish between distinct – and differently abled – categories of moral agent. When it comes to the prospect of synthetic moral agents, we have reason to assume that they would not simply replicate our human example. However speculative in our current context, it would be naive to rely on their slotting effortlessly into the normative frameworks that we have constructed. The result could be catastrophic. Regardless of how powerfully we embrace, and how passionately we defend, particular moral responsibilities in world politics – such as responsibilities to exercise restraint in war – these commitments become meaningless (however well intentioned) if responsibilities are assigned to entities that qualify as moral agents, yet are unable to discharge them. Recognition of the moral

[87] Walzer, *Just and Unjust Wars*, p. 311.

agency of, *inter alia*, the soldier, the citizen, and the state – and what it would mean to one day create a synthetic counterpart – is essential as our 'black box society' encroaches onto the battlefield, into the war room, and (we should assume) upon every domain of international politics.[88]

Finally, the analysis here gestures towards a pressing area for future enquiry. These cases of misplaced responsibility point to a previously unacknowledged source of international norm decay. In the scenarios described above, the strength of the arguments supporting norms of restraint endure, but there is a misalignment between the body we expect to discharge a duty of forbearance and what that body is capable of doing. In the first case of misplaced responsibility, the flesh-and-blood and institutional moral agents to which particular expectations are legitimately directed see themselves displaced as the relevant decision makers. In the second case, we prospectively assign synthetic moral agents responsibilities that would be out of step with their capacities and limitations. How our flawed assumptions about emerging AI-enabled entities could affect, and perhaps erode, currently settled norms in world politics – with potentially far-reaching geopolitical implications – is an important question.

Video Abstract. To view the online video abstract, please visit: https://doi.org/10.1017/S0260210524000202.

Acknowledgements. I am grateful for the following opportunities to present earlier iterations of this argument: at the Leverhulme Centre for the Future of Intelligence (CFI) 'Artificial Agency & Collective Intelligence' Workshop, University of Cambridge, 18 September 2017; as a Shedden Lecture and special seminar for the Australian Department of Defence, 28 November 2018 and 31 January 2019; to the Centre for Moral, Social and Political Philosophy Seminar Series, ANU, 4 March 2019; at the 'Security in Society 5.0' Symposium, Keio University, 23 April 2019; as a public lecture at Tokyo University, 6 December 2019; at the International Studies Association annual convention (online) 2021; as a guest lecture (online) to the Center for War Studies Research Colloquium at the University of Southern Denmark (USD), 4 November 2021; at the Global Governance Colloquium WZB Berlin Social Science Center, 30 June 2022; and at the European International Studies Association (EISA) conference in Potsdam, 8 September 2023. I am grateful to the CFI Workshop participants – Joanna Bryson, Sean Fleming, Christian List, Onora O'Neill, Avia Pasternak, Philip Pettit, Huw Price, David Runciman, and Lauren Wilcox – and the Defence, ANU, Keio University, Tokyo University, ISA, USD, WZB, and EISA audiences for their constructive feedback. I would also like to thank Peter Balint, Christian Barry, Lindsay Clarke, Neta Crawford, Ned Dobos, John Dryzek, Arisa Ema, Liane Hartnett, Emily Hitchman, Tuukka Kaikkonen, Cian O'Driscoll, Umut Ozguc, Jonathan Pickering, Mitja Sienknecht, Sheena Smith, Nicholas Southwood, Ana Tanasoca, Xueyin Zha, Michael Zürn, and this journal's anonymous reviewers for incisive written comments on previous drafts, and Bianca Baggiarini, Claire Benn, Tony Burke, Jenny Davis, Thomas Gehring, Bob Goodin, Sarah Logan, Susan Park, Johanna Seibt, and Toby Walsh for helpful discussions of particular points.

Funding statement. The initial stage of this research was supported by a research grant from Google.org and the Association for Pacific Rim Universities (APRU); research towards the final section on 'misplaced responsibility' in war was supported by a Strategic Policy Grant from the Australian Department of Defence.

Toni Erskine is Professor of International Politics in the Coral Bell School of Asia Pacific Affairs at the Australian National University (ANU) and Associate Fellow of the Leverhulme Centre for the Future of Intelligence at Cambridge University. She is Chief Investigator of a two-year research project on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making' funded by the Australian Department of Defence and serves as Academic Lead for the United Nations Economic and Social Commission for Asia and the Pacific (UN ESCAP)/Association of Pacific Rim Universities (APRU) 'AI for the Social Good' Research Project. She recently served as Director of the Coral Bell School at ANU (2018–23) and Editor of *International Theory: A Journal of International Politics, Law, and Philosophy* (2019–23). Her research interests include: the moral agency and responsibility of formal organisations in world politics; the ethics of war; cosmopolitan theories and their critics; joint purposive action and informal associations in the context of global crises; the responsibility to protect populations from mass atrocity crimes (R2P); the impact of artificial intelligence (AI) on organised violence; and normative IR theory. She is the recipient of the International Studies Association's 2024 International Ethics Distinguished Scholar Award.

[88]The notion of a 'black box society' is taken from Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2016).

Cite this article: Toni Erskine, 'AI and the future of IR: Disentangling flesh-and-blood, institutional, and synthetic moral agency in world politics', *Review of International Studies*, 50 (2024), pp. 534–559. https://doi.org/10.1017/S0260210524000202