Roberts, J.T.F. and Quigley, M. 'Being Novel? Regulating Emerging Technologies Under Conditions of Uncertainty' In Lawrence, David and Morley, Sarah (Eds) *Novel Beings: Regulatory Approaches for a Future of New Intelligent Life* (Edward Elgar Publishing, 2022) pp.140-170

Being Novel? Regulating Emerging Technologies Under Conditions of Uncertainty*

Abstract

If we accept that at some point novel beings will be brought into existence, then we need to consider how the law should take account of (the emergence of) such beings. This is a sticky problem, because any attempt to engage in preparatory regulation with respect to novel beings is mired in uncertainty. Put simply, we do not know what type of beings they will be, either in terms of their physical nature/embodiment or mental/cognitive characteristics. As a consequence, we lack the relevant context-dependent information needed to propose a detailed regulatory regime. In light of this epistemic uncertainty, in this chapter we do not propose a detailed account of law and regulation for novel beings. Instead, we outline a range of normative principles which could help guide the regulation of precursor technologies without undermining our ability to appropriately regulate emerging novel beings in the future.

Introduction

If we accept that at some point novel beings - be they synthetic, biological, or biohybrid in nature – will be brought into existence, then we need to consider how the law should take account of (the emergence of) such beings. This is a sticky problem, because any attempt to engage in preparatory regulation with respect to novel beings is mired in uncertainty. Principally this uncertainty relates to the fact that there is a large - perhaps insurmountable -epistemic gap when it comes to all manner of (legally and morally) significant facts about novel beings. Put simply, we do not know what type of beings they will be, either in terms of their physical nature/embodiment, mental/cognitive characteristics, or how these will influence their preferences and values.[1] Essentially this means that, as these beings do not yet exist, we do not have access to the relevant context-dependent information needed to propose a detailed regulatory regime.

In light of these epistemic difficulties and uncertainties, in this chapter we do not propose a fine-grained account of law and regulation for novel beings. Instead, we outline some tentative normative principles which could help guide the regulation and governance of both novel beings, once

* This is the author accepted manuscript version of our chapter. Our thanks to the Editors, Laura Downey, and the anonymous reviewer for their comments on this piece. Any errors or omissions remain our own. Work on this was generously supported by a Wellcome Trust Investigator Award in Humanities and Social Sciences 2019-2024 (Grant No: 212507/Z/18/Z).

[1] Alex McKeown, 'What Do We Owe Novel Synthetic Beings and How Can We Be Sure?' (2021) 30 Camb Q Healthc Ethics 479.

they come into existence, and precursor technologies. By precursor technologies, we mean those technologies which exist now, but which are likely to serve as important stepping- stones to the eventual emergence of particular novel beings.

In this chapter we will focus primarily on one type of precursor technology: digital-based task-specific expert systems. These are computer programmes consisting of a knowledge base and a set of rules for applying this knowledge (known as an inference engine) which are designed to perform specific tasks.[2] As such, they represent potential precursors to the kinds of advanced artificial general intelligences (AGIs) which would qualify as novel beings possessing moral status. Although we focus primarily on AGI's to illustrate how the principles can be applied to both precursor technologies and novel beings, the principles outlined are intended to apply to other forms of precursor technologies (and their associated novel beings) as well. It may be that, upon further examination, some of the principles may require modification or re-interpretation to govern the development of other forms of precursor technologies and their associated novel beings. Given space-constraints, however, we cannot explore the specific application of the principles outlined to other forms of novel being.

In order to think about what principles might be useful and appropriate, we first set out why the emergence of novel beings might be a problem for the law. Given the multiple levels of uncertainty we identify, we then move on to discuss different potential options in 'regulating for uncertainty'.[3] Our intermediate conclusion is that, although some form of anticipatory regulation is required, at least initially the epistemic uncertainty is such that a principles-based approach may be the most fruitful for both regulators and regulatees. And, from this, more specific and granular law, regulation, and governance may develop once some of the uncertainties regarding these (potential) novel beings have been resolved. Principles based approaches, we argue, allows us to engage in anticipatory regulation and governance of precursor technologies without hampering our future ability to regulate the existence of novel beings (should they emerge). However, in so doing, we note some (not insignificant) drawbacks of such an approach. In the last substantive section of the chapter, we suggest and outline four principles which could help to guide law and regulation regarding (the emergence of) novel beings. These are the principles of: non-domination, responsibility, explicability, and non-harm.

---

[2] David Lawrence and Sarah Morley, 'Regulating the Tyrell Corporation: The Emergence of Novel Beings' (2021) 30 Camb Q Healthc Ethics 421, 423.

[3] Nayha Sethi, 'Regulating for Uncertainty: bridging blurred boundaries in medical innovation, research and treatment' (2019) 11 LIT 112. For a general overview of some of the issues and potentials regarding the regulation of (emerging) technologies see Lyria Bennett Moses, 'Regulating in the Face of Sociotechnical Change' in Roger Brownsword, Eloise Scotford, and Karen Yeung (eds) *The Oxford Handbook of Law, Regulation, and Technology* (OUP 2018).

We end the chapter by outlining some drawbacks of our suggested approach, noting that, in the longer term, it is likely that a multi-layered (and somewhat hierarchical) approach encompassing different sorts of regulatory and governance strategies and tools may be necessary.

2. The Emergence of Novel Beings?

Despite the claim inherent in the name, the creation of 'novel beings' is not entirely new. Humans have been crossbreeding both plants and animals to create new varieties and species for a very long time.[4] Humans also have a long history of selectively breeding individuals within species to select for desired characteristics (e.g. different breeds of dog) and breeding between species to create hybrids (e.g. mules and hinnies). Having said that, the types of novel beings at issue in this book, and of interest in this chapter, are arguably of a different ilk. Whereas new species of plants and animals humans have bred in the past might be viewed as simply tweaks on the existing system over millennia,[5] the forms of novel being we will be discussing in this chapter are distinctive in two ways.

First, they are novel in a stronger sense of the word. Whereas premodern forms of plant and animal breeding were different in some regards to their predecessors, they can be understood as variations on a theme which are continuous with what came before. The novel beings we are talking about in this chapter, however, are likely to be entirely 'new' (in the not existed before sense) types of beings with new forms of consciousness, capable of having their own motivations, goals, and values. The creation of novel beings with the 'capacity for consciousness at the same level as, or surpassing ours, and perhaps meeting the philosophical criteria for philosophical personhood',[6] therefore, involves a qualitative jump not present in the cross-breeding of plants and animals.

Second, prior to the widespread acceptance of genetics in the last 100 years,[7] the creation of new varieties of crops and breeds of animal through selective breeding was a slow and haphazard process.[8] Although humans had a role in selecting for desired traits in the plants and animals they bred, they had little control over the process. In this sense, we could say new types of beings 'emerged' from the process of selective cross-breeding as opposed to them being 'created'. The difference between creation and emergence is that creation implies a greater role for agency and control than emergence does. Whereas describing something as emerging implies that

[4] Noel Kingsbury, *Hybrid: The History & Science of Plant Breeding* (University of Chicago Press 2009).
[5] Dominic Berry, 'Plants are Technologies' in Jon Agar and Jacob Ward (eds), *Histories of Technology, the Environment, and Modern Britain.* (UCL Press 2019)
[6] David Lawrence and Margaret Brazier, 'Legally human? 'Novel beings' and English Law' (2018) 26 Med L Rev 309.
[7] Kingsbury (n 4) 40.
[8] Kingsbury (n 4) 47.

the thing in question is appearing (or recently appeared), the term creation implies the contribution of a creator.[9] The types of novel beings we will be considering may soon come into existence through advances in artificial intelligence (AI), synthetic genomics, gene printing, and/or cognitive enhancement.[10] There are, thus, a variety of different types of novel beings that could emerge and inhabit the earth with us. These could include: artificial general intelligences,[11] genetically modified animals,[12] cognitively enhanced humans,[13] or synthetic biological constructs.[14] It is not completely clear yet *whether* we will be able to create the types of beings we will be considering in this chapter. Neither is it completely clear *when* they will come in to existence if their creation proves to be technically feasible. In debates regarding AI, for example, it is a contentious issue whether the creation of artificial general intelligence is physically possible; that is, whether an AI which could think and learn like us is compatible with the laws of nature.[15] And even if it is physically possible, there are still enormous challenges to be overcome before an artificial general intelligence is technologically possible.[16] These challenges include, amongst others, the improved ability to recognise visual inputs[17] and the ability to accurately model language, creativity, and human emotion.[18] Part of the difficulty in doing these things is that we do not yet fully understand how

[9] Given that control is a matter of degree, there is a spectrum between completely random emergence (where novelty appears without agency) and fully intentional and deliberate creation (where agency directs the process entirely). What is distinctive about the kinds of novel beings we consider is the greater extent to which we have control over the process of bringing them into existence, even if the process is not complete. They are thus closer to the 'creation' side of the spectrum than pre-genetic plant breeding and animal husbandry practices. However, as we intend our discussion to be applicable to scenarios in which novel beings are brought into existence in the absence of full control over the process, we use the term 'emergence' to describe the appearance of novel beings as opposed to the term creation.

[10] David Lawrence, 'Commentary: On Understanding Novel Minds' (2019) 28 Camb Q Healthc Ethics 599; David Lawrence, 'Advanced bioscience and AI: debugging the future of life' (2019) 3 Emerging Top Life Sci 747; David Lawrence and Sarah Morley 'Novel Beings: Moral Status and Regulation' (2021) 30 Camb Q Healthc Ethics 415*;* Lawrence and Morley, 'Regulating the Tyrell Corporation' (n 2) 421.

[11] Lawrence and Brazier (n 6) 314; Lawrence and Morley (n 2) 423.

[12] Gardar Arnason, 'The Moral Status of Cognitively Enhanced Monkeys and Other Novel Beings' (2021) 30 Camb Q Healthc Ethics 492

[13] Lawrence and Brazier (n 6) 312; Lawrence and Morley (n 2) 422.

[14] Lawrence and Brazier (n 6) 316.

[15] Paul Churchland and Patricia Smith Churchland, 'Could a Machine Think?' (1990) 262 Sci Am 32; Kevin Warwick, *Artificial Intelligence: The Basics* (Routledge 2012) 65; Joanna Bryson, Mihailis Diamantis, and Thomas Grant, 'Of, for, and by the people: the legal lacuna of synthetic persons' (2017) 25 A.I.& L 283; Margaret Boden, *Artificial Intelligence: A Very Short Introduction* (OUP 2018) 130.

[16] Boden (n 15) 130; Lawrence and Morley 'Regulating the Tyrell Corporation' (n 2) 423.

[17] Boden (n 15) 37.

[18] Ibid. 50.

these processes work in humans,[19] making modelling them in an artificial environment extremely difficult.

Despite these difficulties, for the purpose of this chapter, we are going to presume that at some point these kinds of novel beings could be brought into existence. This includes both presuming that their creation might be technically feasible and that doing so would not be explicitly legally prohibited (neither of which might turn out to be the case at some point in the future). If we accept this possibility, then we need to consider how the law could and should take account of such beings. Broadly, their potential future existence raises two types of questions for the law. On the one hand, we have questions of how to regulate *the emergence of* such beings. On the other hand, we have questions about how these beings should be regulated *once they have come into existence*. Each of these limbs of the dilemma raises its own distinct set of questions.

In relation to the *emergence* of novel beings, for example, questions arise about whether we should allow the creation of novel beings at all. Further questions stem from answering this either in the negative or positive. For instance, if we ought not to allow them to come into existence, how can we prevent this from occurring? Would legal prohibition be effective? Or, if we ought to allow them to be created, then how should we regulate their creation? These are difficult questions with no immediately obvious answers. Similarly, those in relation to regulating novel beings once they come into existence seem no easier to resolve. For instance, there will be questions about how they ought to be treated, morally- and legally-speaking, as well as ones about the legal rules needed to govern their behaviour and actions. In setting out some of these, we also note that novel beings will not emerge in a vacuum. They will be preceded, as noted earlier, by precursor technologies. As such, to add to the challenge, there are further questions regarding how we could and should regulate precursor technologies to achieve the goal of either preventing (or facilitating) the development/emergence of novel beings.

For the sceptics out there (and to some degree that includes the authors of this chapter), it is worth noting that even though there are doubts about whether novel beings will arise (and if so, when), there is still value in thinking about how we ought to govern their emergence. The first reason it is important is that, so long as there is a possibility that novel beings will exist, we ought to consider how to respond in order to be prepared for their emergence if it does happen. Of course, in hindsight, we may come to realise that this turned out to be wasted effort if novel beings deserving of moral status are never created; either because we opt to legally prohibit their creation or because they are impossible in principle for a reason we do not yet understand. That, however, does not necessarily mean that we ought not to devote intellectual resources to considering the problem now. The

[19] Ibid. 59.

second reason is that considering these questions before they are actual problems can influence *the development of* the precursor technologies that might give rise to the creation of novel beings.[20] Considering these questions prior to the emergence of the novel beings is valuable even if novel beings are never brought into existence, because of the fact it can help us consider and mitigate problems with the technology we have in the present.

So having briefly made the case for why we ought to consider the legal, regulatory, and governance implications of (the emergence of) novel beings, we freely admit that we could not even begin to answer the questions set out above in this chapter. Such a task would be incredibly complex and daunting. Instead below we identify uncertainty as being one of the principal challenges in this area and outline the layered and multi-level nature of this uncertainty. We then examine some ways in which we could approach the regulation and governance of novel beings, focusing on the example of digital-based task-specific expert systems as potential precursors to artificial general intelligences (AGIs).

3. Regulating for Uncertainty?[21]

As noted earlier, regulating (the development and emergence of) novel beings is a sticky problem for the law, because all attempts at engaging in preparatory regulation inescapably involve high levels of uncertainty. Put simply, we are ignorant of many of the legally (and morally) relevant facts that we would need to know to answer the kinds of questions set out in the previous section. The epistemic gaps in relation to novel beings are multi-layered and encompass different levels of uncertainty. These uncertainties include at least four aspects. First, as noted earlier, we do not yet know *whether* such novel beings will come to exist, precisely *how* their existence could be brought about, or *when* they will be brought into existence. Second, we do not know, beyond informed speculation, which forms of precursor technologies we are currently developing (if any) will ultimately be the successful springboards for the emergence of a novel being. Third, we have little knowledge regarding what the impacts of bringing these beings into existence will be.[22] Fourth, and potentially most importantly, we do not yet have a clear idea of what these novel beings will be *like*. We cannot anticipate what their physical make-up will be, what kinds of cognitive abilities they will have, what their mental lives will be like phenomenologically (i.e. from the inside) or, as McKeown has put it, "how

[20] Deborah G Johnson, 'Software Agents, Anticipatory Ethics, and Accountability' in Gary E Marchant, Braden R Allenby, and Joseph R Herkert (eds), *The Growing Gap Between Emerging Technologies and Legal Ethical Oversight: The Pacing Problem* (Springer 2011) 66; Virginia Dignum, 'Responsibility and Artificial Intelligence' in Markus D Dubber, Frank Pasquale, and Sunit Das (Eds), *Oxford Handbook of Ethics of AI* (OUP 2020) 221

[21] Sethi (n 3) 112.

[22] Our thanks to the anonymous reviewer for prompting us to include this.

their corporeality shapes their options, preferences, values, and is constitutive of their moral universe".[23] The corollary to all of this is that novel beings represent what Giddens terms manufactured risk.

> Manufactured risk is risk created by the very progression of human development, especially by the progression of science and technology. Manufactured risk refers to new risk environments for which history provides us with very little previous experience. We often don't really know what the risks are, let along how to calculate them accurately in terms of risk.[24]

In short, as these beings do not yet exist, we do not yet have access to the relevant context-dependent information needed to propose a detailed regulatory and/or governance regime to govern their existence.

Given this level of epistemic uncertainty, what is to be done? We could take a 'wait and see' approach.[25] And in some ways this might appear to be a better route, as at least some of the uncertainties would resolve themselves if we simply waited. Indeed, this is frequently the approach of the law when it comes to new and emerging technologies.[26] We often do not know what the full implications of a new form of technology will be before or even while it is in development. As a consequence, the law can be reactive.[27] As Brownsword notes, in the main, "regulators do not try to anticipate the development of a new technology; rather, they address new technologies only when they must".[28] They do not take a pre-emptory approach and tend, at least in the early stages (until they get a better sense of the risks and benefits), to draw on existing legal and regulatory regimes/elements and adapt them to the new technologies.[29] An example of such an approach discussed by Fincke is the current approach to smart contracts.[30] Regarding these, the Law Commission recently concluded that:

> Current legal principles can apply to smart legal contracts in much the same way as they do to traditional contracts, albeit with an incremental and principled development of the common law in specific contexts. In general, difficulties associated with applying the existing law to smart legal

[23] McKeown (n 1) 479.
[24] Anthony Giddens, 'Risk and Responsibility' (1999) 62 MLR 1, 4.
[25] For a good discussion of the 'wait and see' approach generally (as well as specifically in the context of blockchain technologies) see Michèle Finck, Blockchain Regulation and Governance in Europe (CUP 2018) 154-155.
[26] Ibid.
[27] Roger Brownsword, 'Legal Regulation of Technology: Supporting Innovation, Managing Risk, & Respecting Values' in T L Pittinsky (ed), *Science, Technology, & Society: New Perspectives & Directions* (CUP 2019) 109; Lawrence and Morley (n 10) 415.
[28] Brownsword (n 27) 109.
[29] Ibid. 110.
[30] Fincke (n 25) 154.

> contracts are not unique to them, and could equally arise in the context of traditional contracts.[31]

Arguably only time will tell whether this conclusion will hold regarding this particular technology.

In the case of novel beings, this 'wait and see' approach would seem, at least on the face of it, not to be satisfactory. This is because the consequences of not acting/pre-empting/preparing could be substantial. The problem is neatly encapsulated by the so-called 'Collingridge dilemma' (also known as the dilemma of control)[32] which, as articulated by Bennett Moses, holds that "attempts to control a technology early in its development suffer from the difficulty of not knowing its final form and ultimate effects, while attempts to control a technology after it has become entrenched are virtually impossible".[33] We have already set out the epistemic uncertainties with regards to novel beings, but the message here is that if we wait until they emerge or are near emergence, it will almost certainly be too late to prevent this (if that is what is decided) or to do much to influence the shape and trajectory of these beings' lives (if they ought to be permitted to come into existence in the first place). As such, despite the fact that the uncertainties in relation to novel beings seem deeper than those surrounding other emerging technologies, engaging in some form of anticipatory governance and/or preparatory regulation is likely to be necessary.[34] We may find in the future that our attempts at pre-empting the technologies focused on possibilities that failed to materialise, but nevertheless such attempts are still valuable, if only because they can help to guide the regulation of precursor technologies.

So does this mean that we ought to go ahead and devise a detailed new legal regime? Well this is also not the solution in this case. Bennett Moses notes four things about technological change and the law: first, that it can bring about "pressure to enact new laws"; second, there is often "a need to resolve uncertainties in the application of law in new contexts";

---

[31] Law Commission, *Smart Legal Contracts: Summary* (LC 401, 2021) https://www.lawcom.gov.uk/project/smart-contracts/ accessed 7 April 2022.

[32] David Collingridge, *The Social Control of Technology*. (Frances Pinter 1980) 19.

[33] Lyria Bennett Moses, 'Sui Generis Rules' in Marchant, Allenby and Herkert (n 20) 89. See also Graeme Laurie, Shawn HE Harmon, and Fabiana Arzuaga, 'Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty' (2012) 4 LIT 1, 5-6. For a recent analysis of neglected aspects of Collingridge's work in the context of responsible research innovation see Audley Genusa and Andy Stirling, 'Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation' (2018) 47 Res Policy 61.

[34] For some work on anticipatory governance generally, as well as in relation to emerging technologies broadly see Laurie et al (n 33), Leon S Fuerth, 'Foresight and Anticipatory Governance' (2009) 11 Foresight 14, David H Guston, 'Understanding "Anticipatory Governance" (2014) 44 Soc Stud Sci 218, David Sarewtiz, 'Anticipatory Governance of Emerging Technologies' in G.E. Marchant et al. (eds.), *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem* (Springer 2011).

third, existing "legal rules … apply poorly in [these] new contexts"; and fourth, existing law can become obsolete in the face of new developments.[35] The consequence of this, she maintains, can be a tendency to create *sui generis* legal rules – that is a set of particularised rules - to deal with new technologies or technological change. Yet this may not always be the best route. Laurie and colleagues give the example of the ban in cloning in the Human Fertilisation and Embryology Act 1990 as one of "legislating too early" in the face of uncertainty.[36] As originally enacted, section 3(3)(d) prohibited the granting of a licence for "replacing a nucleus of a cell of *an embryo* with a nucleus taken from a cell of any person, embryo or subsequent development of an embryo".[37] However, as time went on it became clear that this wording only captured one form of 'cloning'. At the time of drafting scientific advances relating to stem cell production from so-called therapeutic cloning via cell nuclear replacement (whereby the nucleus of an oocyte is replaced with that of another cell) had not been anticipated.[38]

In the case of novel beings which we are considering here, acting too early could result in law which is quickly outpaced by technological developments or in advances that were simply never envisaged at the time of drafting of the relevant statutes or regulations. The reason is that the epistemic uncertainties *right now* are so great that we do not have enough context-specific information to apply existing legal rules, let alone create any kind of new *sui generis* regime. This is quite apart from the fact that acting too early to create a *sui generis* regime for each and every potential new novel being and its technologies would be costly, time-consuming, and impractical.

Given this, if we ought not to simply 'wait and see' and we are not in a position to either apply existing law or create bespoke legal frameworks, what is to be done?

4. A Principles-Based Approach

Our suggestion is that, in light of the epistemic difficulties we face trying to understand *what* novel beings will be like and *how* they will emerge, a principles-based approach might prove useful *at the moment*. As we get closer to making novel beings a reality, our ignorance in these respects will likely diminish, enabling us to move beyond guiding principles. However,

[35] Ibid. 77.
[36] Laurie et al (n 33) 6, note 23.
[37] Emphasis added.
[38] There is not the space to go into it here, but the result of this was the controversial *R (on the application of Quintavalle) v Secretary of State for Health* [2003] 2 WLR 692 in which the Pro Life Alliance (PLA) argued that the HFEA could not grant a licence for CNR under its remit as CNR fell outside the scope of the 1990 Act. Although the high court found for the at first instance, this was overturned on appeal and the Court of Appeal judgment affirmed in the House of Lords.

until this occurs, a principles-based approach would be a feasible and practical one. So what do we mean by a principles-based approach (PBA)? And what are the potential benefits and drawbacks of such an approach?

4.1 Principles, Rules, Regulation, and Governance

PBAs take their cue from the narrower principles-based-regulation (PBR).[39] As outlined by Black, Hopper, and Band, "[p]rinciples-based regulation means moving away from reliance on detailed, prescriptive rules and relying more on high-level, broadly stated rules or Principles to set the standards by which regulated firms must conduct business".[40] PBR is an approach which has found favour in the United Kingdom in a number of areas, these include stem cell research, finance, and the legal profession.[41] One of the most successful examples of its use, and outlined in detail by Devaney,[42] is the approach of the Human Fertilisation and Embryology Authority (HFEA), which is responsible for the regulation of fertility clinics and the use of human embryos for research.[43] Whilst the HFEA derives its regulatory mandate directly from the Human Fertilisation and Embryology Act 1990 (as amended), the Act itself does not contain all the details of the regulatory regime. Instead the Authority is required by virtue of section 8(1)(ca) to set out a statement of general operating principles which are to guide both the work of the HFEA itself and any activities governed by the Act.[44] In practice this has been achieved by integrating these general principles into the HFEA's codes of conduct.[45]

Currently, the code of conduct contains a set of 13 'regulatory principles' which those carrying out activities licenced under the Act must

---

[39] As will be explained below, we follow Laurie and Sethi in talking about 'principles-based approaches' rather than just principles-based regulation. See Graeme Laurie and Nayha Sethi, 'Towards Principles-Based Approaches to Governance of Health-Related Research Using Personal Data' (2013) 1 EJRR 43, Nayha Sethi and Graeme Laurie, 'Delivering Proportionate Governance in the Era of eHealth: Making Linkage and Privacy Work Together' (2013) 13 Med Law Int 168.

[40] J Black, M Hopper, and C Band, 'Making a success of Principles-based regulation' (2007) 1 LFMR 191.

[41] Although note the criticisms regarding its use in the financial sector, with some noting that weaknesses with this approach contributed to the 2007-08 financial crisis. See, for example: Julia Black 'Forms and Paradoxes of Principles-Based Regulation' (2008) 3 CMLJ 425; Julia Black, 'The Rise, Fall, and Fate of Principles Based Regulation' in Kern Alexander and Niamh Moloney (eds), *Law Reform and Financial Markets* (Edward Elgar 2011); Sarah Devaney, 'Regulate to Innovate: Principles Based-Regulation of Stem Cell Research' (2011) 11 Med L Int. 53.

[42] Devaney (n 41); Sarah Devaney, *Stem Cell Research and Collaborative Regulation of Innovation* (Routledge 2014) Ch. 2.

[43] Human Fertilization & Embryology Authority, 'How we regulate' <https://www.hfea.gov.uk/about-us/how-we-regulate/> accessed 23rd February 2021.

[44] For more on this see: Devaney (n 42) 38-40.

[45] Ibid. 39-40.

adhere to.[46] As noted by Devaney, "[t]hey are intended to be high-level, broadly stated and self-contained provisions containing qualitative terms as opposed to prescriptive rules".[47] Principle 1, for instance, states that licensed centres must "treat prospective and current patients and donors fairly, and ensure that all licensed activities are conducted in a non-discriminatory way".[48] Whilst Principle 2 requires that centres must "have respect for the privacy, confidentiality, dignity, comfort and well-being of prospective and current patients and donors".[49] Such an approach could, on the face of it, be contrasted with more rigid rules-based regulation, which "relies on compliance with specific rules".[50] We say 'on the face of it', because as we will see, attempting to identify a clear distinction between rule-based and principles-based approaches may not prove particularly straightforward or illuminating.[51]

Indeed, as Laurie and Sethi argue, there are numerous instances "whereby so-called 'principles' in legislation are, in fact, operating as rules or in a rule-like manner."[52] They draw on data protection legislation to aptly illustrate this. Whilst the old Data Protection Directive explicitly contained eight 'principles', these did not function as such in practice, because "derogation from any of the principles represent[ed] a clear breach of law."[53] As anticipated by Laurie and Sethi at the time they were writing, the current EU General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) is also constructed around 'principles'.[54] These seven principles are also to be found in the Data Protection Act 2018 (and the post-Brexit UK GDPR).[55] Yet the move from Directive to Regulation actually reduces any leeway in interpretation and application of the 'principles' which EU countries, for instance, may have had in the past.[56] By the same token, the example HFEA principles given above are in reality a mixture of more or

---

[46] Human Fertilisation and Embryology Authority, 'Code of Practice' (9th Edition, HFEA 2021) 12-13. <https://portal.hfea.gov.uk/media/1756/2021-10-26-code-of-practice-2021.pdf > accessed 7th April 2022.
[47] Devaney (n 42) 40.
[48] HFEA (n 46) 12.
[49] Ibid.
[50] Laurie and Sethi (n 39) 45.
[51] Nayha Sethi, 'Rules, Principles, and the Added Value of Best Practice in Health Research Regulation' in Graeme Laurie, Edward Dove, Agomoni Ganguli-Mitra, *et al.* (eds), *The Cambridge Handbook of Health Research Regulation* (CUP 2021) 169.
[52] Laurie and Sethi (n 39) 46.
[53] Ibid. 50.
[54] Ibid. 51. At the tine their article was written and published the EU Data Protection Regulation was still in draft form and had yet to be passed and implemented.
[55] In the EU GDPR these are: Article 5 - Principles relating to processing of personal data; Article 6 - Lawfulness of processing; Article 7 - Conditions for consent; Article 8 - Conditions applicable to child's consent in relation to information society services; Article 9 - Processing of special categories of personal data; Article 10 - Processing of personal data relating to criminal convictions and offences; and Article 11 - Processing which does not require identification. These are mirrored in the 2018 Act and thus the UK GDPR.
[56] Ibid.

less rule-like features, with each Guidance Note within the Code of Practice outlining the mandatory requirements as set out in the HFE Act and HFEA's interpretation of these requirements, along with other less rule-like aspects of the guidance. Given all of this, we agree with Sethi who favours "talk of 'rule and principle-type features'" rather than attempting any definitive delineation between rules and principles.[57]

In a similar vein, attempting to draw a distinction between approaches which could be classified as regulation-based and those which are governance-based may not always be straightforward; and again potentially not all that helpful. Broadly speaking regulation has a narrower focus than governance,[58] with some viewing the former as a subset of the latter.[59] For our purposes, we take the lead again from Laurie and Sethi who distinguish between the two as follows:

> …while *regulation* will tend to be a state-driven, vertically-orientated, top-down, command-and-control deployment of formal (hard law) instruments, *governance* is often a far more horizontally-orientated enterprise, more likely driven by local actors, and more reliant on soft law options such as guidance or professional codes.[60]

What is interesting about the example HFEA principles set out above (and indeed other HFEA principles) is that they ought not to be viewed strictly - or necessarily - as pure 'regulatory' principles (albeit they are labelled as such within the HFEA's Code of Practice[61]). In practice they are, to varying degrees, principles which encapsulate a mixture of regulation and governance features. They can be viewed as regulatory to the extent that they are top-down principles put in place via the state backed regulator. Whilst they are not command and control instruments in and of themselves, the mandate does derive from hard law. However, they can be viewed as tending more towards a governance approach in implementation. This is because how some aspects of these high-level principles are cashed out and implemented is dependent on the specifics of local guidance and practice.

---

[57] Ibid. There is, of course, a large literature on rules and principles, including within legal theory (where the distinction featured prominently in the Dworkin-Hart debate). We do not have the space to delve into this in this chapter, which in any case has been dealt with extensively, and more competently, by others elsewhere. See, for example, Nayha Sethi, 'Reimagining Regulatory Approaches: On the Essential Role of Principles in Health Research Regulation' (2015) 12 Scripted 91; Ronald Dworkin, 'The Model of Rules' (1967) 35 U Chi L Rev 25; H. L. A. Hart, *The Concept of Law* (2nd Edition, OUP 1994) 259; Michael Bayles 'Hart vs. Dworkin' 10 L & Phil 351.

[58] See generally Julia Black, 'Critical Reflections on Regulation' (2002) 27 Aust J Leg Philos 1.

[59] John Braithwaite, Cary Coglianese, and David Levi-Faur, 'Can Regulation and Governance Make a Difference?' (2007) 1 Reg Gov 1, 3.

[60] Laurie and Sethi (n 39) 47.

[61] HFEA (n 46) 13-14.

For Laurie and Sethi, therefore, "it is preferable to talk of a principles-based *approach*".[62] Doing so recognises that (1) distinctions between rules and principles, as well as between regulation and governance, are not always clear-cut and (2) a mix of approaches may be necessary depending on the context. For them, the purpose of principles is to initiate and drive suitably reflective and nuanced conversations about the target behaviours or technologies. Thus regardless of terminology, the important takeaway is this:

> Principles should be seen as fundamental starting points to guide deliberation and action. Thus, their purpose is to point actors or decision-makers in the direction of the relevant values and considerations to be taken into consideration when a particular decision or course of action is being contemplated… Principles require reflection and justification for actions which, in themselves, are signals of 'good governance'.[63]

In this spirit, in section 5, we outline four potential starting principles for the regulation and governance of (the emergence of) novel beings. These are the principles of non-domination, responsibility, explicability, and non-harm. For each of these principles we will outline the higher order proposition which they represent, why we think they might be helpful, and some potential drawbacks. However, before we do this, it is worth taking a closer look at some of the general benefits and disadvantages of a principles-based approach.

4.2 Weighing Up Principles-Based Approaches

It seems that there are a number of epistemic and pragmatic benefits to using a principles-based approach when regulating under conditions of uncertainty. As we have already noted, we do not know enough to do anything more concrete/particularised at this point in time (the first horn of the Collingridge dilemma). Yet we do need to do *something* otherwise by the time novel beings exist - or are on the brink of existence - it will likely be too late to do anything substantive to alter the direction of the technologies (the second horn of the dilemma). Given this, first and foremost, a principles-based approach offers a way to do *something*. But importantly, it gives us a way to do something which can be action-guiding.

Whilst general principles do not give us the content of what must be done, they at least lay out the contours to help us decide on the content. We can see this in the principles set out in the HFEA Code of Conduct. The two examples of the HFEA principles given earlier (fair treatment/non-discrimination and respect for privacy, confidentiality, etc.) do not prescribe or give detail about *how* they ought to be enacted or achieved, but they do contain the normative target to aim for. We accept that there might be

[62] Laurie and Sethi (n 39) 44, Sethi (n 51) 167.
[63] Laurie and Sethi (n 39) 46.

reasonable disagreement, especially in difficult boundary cases, about whether or how a given principle applies to a particular situation. However, not all cases involve grey areas. In more clear-cut cases, having a general statement of the normative goal is sufficient to guide conduct. Secondly, and relatedly, in setting out the normative contours rather than detailed content, a principles-based approach offers a degree of flexibility.[64] There could be multiple ways in which the principles could be enacted and implemented, allowing us to correct course or tweak details as we go along. As such, we obviate the need to specify too early the exact shape of the regulatory and governance space which will flow from the principles, allowing it to develop organically from the application and re-interpretation of the general principles. As Laurie and Sethi note, it negates the need for "detailed anticipatory drafting for every perceivable situation";[65] something which in any case would be nigh on practically impossible.

This is not to say that a principles-based approach is a panacea. Indeed some of the purported strengths of this bring their own challenges. For instance, a significant problem with generality and flexibility as strengths is that – depending on how they are cashed out - they can become weaknesses. To wit, if principles are too general they may not be substantively action-guiding. Alternatively (or indeed in addition), the principles could be open to interpretations which are simply too wide; in which case they may be action-guiding, but permit any number of actions which were not intended at the outset or which actually run counter to the original underlying regulatory aims. The challenge is to ensure that regulation is flexible enough to be adequately responsive, without losing too much transparency or specificity. Moreover, as Brownsword notes, "[f]lexibility comes at the price of predictability".[66] A lack of predictability is problematic under any particular regulatory regime, because regulatees need to know what their obligations are and what they must do to fulfil those obligations.

We will come back to these potential difficulties after we have discussed our potential principles for the regulation and/or governance of (the emergence of) novel beings. But for now we simply want to note that there are ways to mitigate, or perhaps avoid entirely, the sorts of negative sequelae that can attend a principles-based approach. The first is that each principle could be accompanied by a fuller explanation of the purpose and normative aim of that principle. That way when it comes to filling in the detail at a later date regulators (and regulatees) have a meaningful reference point to help with appropriate interpretation.

Second, and relatedly, an account, not just of the purpose of each individual principle, but of the principles as a collective could be given. This could help to reduce ambiguity over the normative direction of travel

[64] Black, Hopper, and Band (n 40) 193.
[65] Laurie and Sethi (n 39) 45.
[66] Brownsword (n 27) 127.

intended. For instance, the HFEA Code of Practice states that "[t]he Code of Practice contains regulatory principles for licensed centres, and guidance notes which provides guidance to help clinics deliver *safe, effective and legally compliant treatment and research*."[67] This thus seems to be a statement of intent regarding the normative purposes of the principles in the code of practice; that is, the normative goal is the delivery of 'safe, effective, and legally compliant treatment and research'.

Third, as noted by Devaney, principles-based approaches can be used in conjunction with more prescriptive rules.[68] These rules can be used to set boundaries which must not be crossed or to set out immoveable lines. For example, whilst the HFEA Code of Practice contains the overarching principles and states that "[t]he regulatory principles inform every part of this Code of Practice",[69] the Code is also very clear on which elements constitute mandatory requirements as laid down either in the HFE Act (as amended) or some other aspect of law. For instance, Principle 1 requires that patients and donors and treated fairly and in a non-discriminatory way.[70] But this is then detailed in terms of obligations under, amongst other things, the Equality Act 2010; noting that a fertility centre's policies and guidance must comply with Act's requirements. Another example of principles being used in conjunction with more prescriptive elements can be found in the General Medical Council's *Good Medical Practice*. Here the guidance often contains sub-principles which make the guidance more specific. For instance, para 14 says "you must recognize and work within the limits of your competence".[71] Then, as a specification, 14.1 says "you must have the necessary knowledge of the English language to provide a good standard of practice and care in the UK".[72] This makes clear that speaking English is a part of what makes a practitioner working in the UK competent to practice medicine. These more prescriptive elements help practitioners interpret the principles with a view to increasing compliance.[73] Importantly, however, and related to Laurie and Sethi's position that principles should be "starting points to guide deliberation and action",[74] the HFEA very explicitly states that it does not purport to offer the definitive interpretation of the law, and thus application of the principles, within its guidance notes.[75]

---

[67] HFEA (n 46) 11.
[68] Devaney (n 41) 60.
[69] HFEA (n 46) 13.
[70] Ibid.
[71] General Medical Council 'Good Medical Practice' (GMC 2013) 7 https://www.gmc-uk.org/ethical-guidance/ethical-guidance-for-doctors/good-medical-practice> accessed 6th April 2022.
[72] Ibid.
[73] Devaney (n 41) 60.
[74] Laurie and Sethi (n 39) 46.
[75] HFEA (n 46) 11.

Fourth, as recently proposed by Sethi in the context of global health emergencies, principles and any attendant guidance could be complemented by a set of best practice examples. These would function as "a mid-level translational mechanism, serving as a bridge from *text* to *context*, more specific than principle-like guidelines, yet not as specific as rule-like guidelines."[76] Whilst accepting that 'best practice' need not (and is unlikely to) mean ideal practice, Sethi locates the utility of such an approach in its ground-up nature.[77] Best practice examples will come from a range of actors and stakeholders on the ground and will "genuinely reflect the experiences of those involved".[78] Examples of best practice case studies can be found, for instance, on the GMC website, where there are a number of structured vignettes to accompany *Good Medical Practice* and other GMC guidance. These involve reading through a clinical interaction and choosing the appropriate next step. There is then an explanation of what the doctor did (i.e. what the GMC views as 'best practice'). Thus the GMC's scenarios involve not only setting out an example of best practice, but are structured to encourage reflection and learning on the part of the doctors using the resource.[79]

We do not purport that the above suggestions are the only (or even the best) ways to mitigate some of the challenges and increase the utility of a principles-based approach. However, they at least offer a partial solution to some of the drawbacks of this type of regulation. Having set out the stall in this respect, let us now move on to look at four principles which might be helpful in the regulation and governance of (the emergence of) novel beings.

5. Tentative Principles for the Regulation of Novel Beings

In this section we propose four tentative principles that could guide the regulation of both precursor technologies and novel beings (if and when they emerge). These are the principles of non-domination, responsibility, explicability, and non-harm. The principles of responsibility, explicability, and non-harm will be familiar to readers interested in the regulation and governance of AI. The originality of our approach lies in extending this framework by adding a principle of non-domination, and in proposing that these principles should be applied, not just to the regulation of current and future AI, but also to the regulation of other precursor technologies and the novel beings with moral status they might give rise to.

---

[76] Nayha Sethi, 'Research and Global Heath Emergencies: On the Essential Role of Best Practice' (2018) 11 Public Health Ethics 237, 242.
[77] Ibid. 245.
[78] Ibid. Sethi's discussion in this respect focused on those involved in dealing with the global health emergencies such as the H5N1, Zika, and Ebola outbreaks. For a discussion of the use of principles and best practice in the context of the linkage of health data in Scotland see Sethi and Laurie (n 33).
[79] General Medical Council 'Good Medical Practice in Action' <https://www.gmc-uk.org/gmpinaction/> accessed 6th April 2022.

Given the epistemic uncertainties we have outlined, the goal is not to provide a detailed regulatory regime. Instead, the goal of this section is to illustrate how the use of principles could help us guide the regulation of precursor technologies without hampering our ability to regulate novel beings in the future. For each proposed principle we outline what is meant by the principle, how it could be applied to both pre- and post-emergence scenarios, and some of the uncertainties that need to be resolved in each case before the principles could be translated into a detailed regulatory regime to govern (the emergence of) novel beings. The section concludes by arguing that, although there are uncertainties to be resolved before the principles outlined can be applied, a principles-based approach offers us the normative guidance necessary to implement the principles as and when the knowledge required to do so becomes available.

5.1 Non-domination

The first proposed principle is the principle of non-domination. Domination is a relationship that exists between an 'agent' of domination and a 'subject' of domination in which the agent of domination has the power to arbitrarily interfere with the lives of the subject of domination.[80] Following Lovett we understand arbitrariness as being a matter of 'the absence of established and commonly known rules, law or conventions effectively governing the use of power in a social relationship'[81] In other words, domination 'consists in a state of subordination or subjugation, whereby somebody's latitude-to-Φ is dependent on the tolerance or leniency of someone else'[82] (where Φ denotes some action or other).

The principle of non-domination holds that no moral agent ought to be dominated by any other moral agent. Ensuring non-domination is important because not being subject to the arbitrary will of another is an important part of being free. Freedom as non-domination is not focused on minimising interference with any given agent's activities, it is primarily about ensuring that certain agents do not have *arbitrary* power over others. Ensuring non-domination for all requires constraining the power of *all* agents to ensure *no* agent can interfere arbitrarily with *any* other agent. It follows that enforcing the principle of non-domination (and the other principles outlined below) will sometimes require interference with some agents' pursuit of their plans. Importantly, however, this does not constitute domination if the interference is constrained through commonly known rules (such as those enforced by publicly legitimised principles-based regulation).

---

[80] Frank Lovett, 'Domination: A Preliminary Analysis' (2001) 84 Monist 98.
[81] Ibid. 103.
[82] Matthew Kramer, 'Liberty and Domination' in Cecile Laborde and John Maynor (eds), *Republicanism and Political Theory* (Blackwell 2008) 32.

The use of precursor technologies - such as task-specific expert systems described in the introduction - have the potential for domination. Currently this type of AI is largely being developed and deployed by profit-seeking corporations to a variety of business and industry ends. Moreover, as Lawrence and Morley point out, they are doing so in a largely unregulated fashion as these activities are not effectively constrained by existing (UK) companies law.[83] If we allow corporations to continue to self-regulate without any form of public control, the ways in which this technology is deployed will not be subject to commonly known rules that effectively constrain the power of these corporations. Under a pure self-regulation model, corporations are essentially a law unto themselves.[84]

This is a problem even if we grant, for the sake of argument, that we ought to trust that the commercial goals of these corporations are in line with the wider societal interests (which is a big concession). From a domination perspective, the objection to a solely self-regulation regime is not only that many corporations will prioritise profit over any negative impacts on wider stakeholders or society (although this is likely to be the case). It is also that people's freedom is limited by corporations under a self-regulatory regime. Even if corporations could be trusted to self-regulate appropriately (which is doubtful), without public control over the regulatory regime, consumers of products which rely on AI are potential subjects of domination.

Applied to the pre-emergence scenario, the principle of non-domination thus implies that we need to have some form of public oversight over the way in which expert systems (and other precursor technologies) are deployed if we are to make the rules which govern their use non-arbitrary. Constraining the power of developers of precursor technology will require moving away from regimes of self-regulation and towards regulation by publicly backed regulators. Implementing a publicly backed regulatory regime might prove to be a challenge as it might conflict with the commercial interests of developers. However, the challenge of imposing regulation on a recalcitrant industry is not necessarily insurmountable.

More difficult problems arise when we try to apply the principle of non-domination to a post-emergence scenario. The first difference is the number of potential domination relationships to consider. Whereas in the case of precursor technologies the agent of domination is the corporation, if novel beings with moral status are brought into existence, novel beings themselves could be both subjects and agents of domination. Novel beings could be subjects of domination if, once they emerge, they are kept as property either by corporations or individuals, as the owners would have the power to arbitrarily interfere in the novel being's life. On the other hand, novel beings could also be agents of domination if, once they emerge, we

[83] Lawrence and Morley (n 2) 427.
[84] Ibid. 422.

cannot effectively constrain their behaviour through commonly known and accepted rules. If this were to occur, novel beings could have the power to arbitrarily interfere in the lives of both other novel beings and humans, thereby dominating them.

In order to comply with the principle of domination, the exercise of power in all these potential relationships needs to be made non-arbitrary. As suggested above, one way of doing this is by ensuring public oversight over the rules that govern the relationship between the different parties through principles-based regulation. Here, however, we face even greater challenges to implementation. Given the uncertainty surrounding what novel beings will be like, we do not currently know how we could effectively constrain their behaviour to ensure they do not become agents of domination or, indeed, whether this would be possible after a certain point in their development.[85] The second challenge to implementing the principle of domination to govern the behaviour of novel beings is that, in virtue of their different embodiment, what constitutes domination for humans may not be domination for a novel being which is differently embodied.[86]

To illustrate: Suppose A, a human, is performing some mathematical calculations. B, another human, has the power to interrupt A's activities or thought processes and compel A to perform another action without reason and whenever they like. This seems like a clear case of arbitrary interference and, hence, domination. Now, the fact that A is unable to perform his original actions in virtue of B asking him to do something else constitutes interference because A cannot do both at once. If A is compelled to do B's bidding, A's original activity must be delayed. The problem is this: the inability to perform more than one *conscious* action (or at most a few) at any one time seems to be a feature of human psychology which novel beings may not share.[87] Computer processors, unlike human minds, can perform multiple operations simultaneously (assuming there is sufficient processing power). It is, thus, not obvious that the same sorts of actions that constitute domination of humans will be the same as those that constitute domination of novel beings. Until this uncertainty surrounding what constitutes domination is resolved, it will be unclear how we ought to implement the principle of non-domination in a post-emergence world.

[85] Our thanks to the anonymous reviewer for this latter point.
[86] Roman Yampolskiy and Joshua Fox, 'Safety Engineering for Artificial General Intelligence' (2013) 32 Topoi 218.
[87] Rene Marois and Jason Ivanoff, 'Capacity Limits of Information Processing in the Brain' (2005) 9 Trends Cogn Sci 296; Earl K Miller and Timothy J Buschman, 'Working Memory Capacity: Limits on the Bandwidth of Cognition' (2015) 144 Daedus 112; Stanislas Dehaene et al (2017) 'What is Consciousness, and could machines have it?' (2017) 358 Science 486, 489.

5.2 Responsibility

The second proposed principle for the regulation of novel beings is the principle of responsibility. The principle of responsibility holds that it should always be possible to hold some entity legally responsible for the negative consequences that might follow from the development and deployment of novel beings or their precursor technologies.

Although novel beings and their precursor technologies have the potential to deliver large benefits to humanity, as with other forms of technology, these benefits will not come without potential harms.[88] Ensuring that some party can be held legally responsible for harms is an important first step to controlling and mitigating them. This is because attributions of responsibility play two roles. First, attributions of responsibility have a prospective role in that they tell us who is required to take action to reduce the likelihood of harms materialising. Second, attributions of responsibility have a retrospective role in that they inform us who is liable for remedial action should the harms occur.[89]

In the context of precursor technologies, responsibility for any potential harms falls on the individuals who develop and deploy precursor technologies such as AI.[90] Here, as before, there are challenges to implementation. The first challenge is that many current AI systems operate as non-transparent, non-interrogable black-boxes, making it hard to understand the cause of a fault (should one occur).[91] This problem is compounded by the complexity of the AI development process. AI systems (and other precursor technologies) are developed by large teams,[92] over an extended period of time, and will often require the use of inputs (e.g. training data sets) developed by other teams working for other

---

[88] Dignum (n 20) 215.

[89] Peter Cane, *Responsibility in Law and Morality* (Hart 2002) 31; HLA Hart, *Punishment and Responsibility* (OUP 2008) 212-215; Karen Yeung, 'Responsibility and AI' (Council of Europe, Study DGI(2019)05) 48. <https://rm.coe.int/responsability-and-ai-en/168097d9c5> accessed 14th July 2021.

[90] Institute of Electrical and Electronics Engineers, 'Ethically Aligned Design: A Vision for Prioritising Human Wellbeing with Autonomous and Intelligent Systems' <https://ethicsinaction.ieee.org> accessed 14th July 2021; Wulf Loh and Janina Loh 'Autonomy and Responsibility in Hybrid Systems' in Patrick Lin, Keith Abney and Ryan Jenkins (eds), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (OUP 2017) 46.

[91] Gunkel, David J. 'Mind the gap: responsible robotics and the problem of responsibility' (2020) 22 Ethics Inf Technol 317; Mark Coeckelbergh 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (2020) 26 Sci Eng Ethics 2057.

[92] David Leslie, 'Understanding Artificial Intelligence Ethics and Safety' (The Alan Turing Institute 2019) 35. <https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety> accessed 20th October 2021; Brent Daniel Mittelstadt et al, 'The ethics of algorithms: Mapping the debate' (2016) Big Data and Society 7.

organisations.[93] Together these features make it difficult to establish who is responsible for what.[94]

In other words, attributing responsibility for the functioning of an AI system is complicated by what is known as 'the problem of many hands'.[95] The problem is that, 'because many different individuals in an organization contribute in many different ways to the decisions and policies, it is difficult even in principle to identify who is responsible for the results.'[96] Individual team members in large organisations will often lack an understanding of (and control over) how their work might be used by other individuals in the organisation, making it difficult to consider them responsible.[97] This lack of control over outcomes is even more acute for organisations that produce component parts or provide services which third parties use to develop AI systems. The problem of many hands thus creates a 'responsibility gap',[98] which can make manufactured risks difficult to mitigate.[99]

There are three main responses to the responsibility gap. The first is to challenge the account of responsibility it relies on.[100] The responsibility gap arises only if one holds that control over an outcome is necessary for responsibility.[101] This may well be the case for moral responsibility, but it is less clear that it is the case for legal responsibility. Unlike moral responsibility, which focuses on apportioning moral blame, legal responsibility has traditionally 'been more sensitive to the interests of victims and of society in security of the person and property',[102] allowing for greater flexibility in who to hold responsible.[103] If we do not require control over outcomes for responsibility, we could assign responsibility for the actions of their subordinates to individuals higher up the corporate hierarchy, regardless of whether they themselves were personally involved in making the decisions or taking the actions that led to harm.[104] Implementing this solution, however, presents a challenge. As Lawrence

[93] Coeckelbergh (n 91) 2057.
[94] Our thanks to the anonymous reviewer fro prompting us to expand our comment regarding the development of technologies by large teams and the problem of many hands.
[95] Dennis Thompson, 'Designing Responsibility: The Problem of Many Hands in Complex Organizations' in Jeroen van den Hoven, Seumas Miller and Thomas Pogge (eds), *The Design Turn in Applied Ethics* (OUP 2017).
[96] Ibid. 32.
[97] Ibo van de Poel et al, 'The Problem of Many Hands: Climate Change as an Example' (2012), Sci Eng Ethics 53; Andreas Matthias, 'The responsibility gap: Ascribing Responsibility for the actions of learning automata' (2004) 6 Ethics Inf Technol 175; Mittelstadt (n 92) 10.
[98] Coeckelbergh (n 91) 2055 'Matthias (n 97) 175'; van de Poel et al (n 97) 50.
[99] Giddens (n 24) 4.
[100] Yeung (n 89) 49.
[101] Coeckelbergh (n 91) 2056.
[102] Yeung (n 89) 50.
[103] Ibid.
[104] Jacob Turner, *Robot Rules* (Palgrave 2019) 98; John Danaher 'Robots, law and the retribution gap' (2016) 18 Ethics Inf Technol 307.

and Morley point out, under current UK company law, it is unclear whether a company director's obligation to promote the success of the company 'stretches to include the responsible development of its products'.[105] Even if it does, and a claim can be brought under the Companies Act 2006, these claims are rarely successful in practice.[106]

Another response to the responsibility gap argument is to create better systems for attributing responsibility.[107] As the problem of many hands arises because we do not know 'who did what' in group projects, it can be mitigated by establishing 'a continuous chain of human responsibility across the whole AI delivery workflow'.[108] This solution, however, may also be challenging to implement. Although the approach may be feasible for AI systems developed within an organisation, establishing a continuous chain of responsibility may be harder when inputs developed by other companies are used. In these cases, the chain of responsibility might have to be established contractually.[109] A third response to the responsibility gap might be to do away with fault requirements entirely, making the legal entity that puts the AI on the market strictly liable for harms caused by AI.[110]

Although fine-tuning the details of such an arrangement will be a complex and important task, 'these challenges all point to familiar solutions based in various ways of holding manufacturers liable.'[111] So whilst there are challenges to implementing the principle of responsibility to govern the development of precursor technologies, these do not appear to be insurmountable.[112]

Even stickier problems emerge when we try and implement the principle of responsibility in cases where we have already created novel beings with moral status, especially if we try and hold the novel being itself responsible for its own performance. The problem here is that it is far from clear what it would mean to hold a novel being responsible for their actions and how this could be achieved in practical terms. Given the current levels of uncertainty regarding what novel beings will be like if they emerge, it is unclear how novel beings would, for instance, pay compensation for harms they cause, or how one could punish a novel being for a lack of compliance

---

[105] Lawrence and Morley (n 2) 427.
[106] Ibid.
[107] Leslie (n 92) 23; Joanna Bryson, 'The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation' in Dubber, Pasquale and Das (eds) (n 20) 6; Helen Nissembaum, 'Computing and Accountability' (1994) 37 Communications of the ACM 79; Mittelstadt (n 92) 13.
[108] Leslie (n 92) 26.
[109] Turner (n 104) 106.
[110] Nissembaum (n 107) 79; Turner (n 104) 91; David C Vladeck, 'Machines Without Principals: Liability Rules and Artificial Intelligence' (2014) 89 Washington Law Rev 149; Danaher (n 104) 307.
[111] Trevor N White and Seth D Baum, 'Liability for Present and Future Robotics Technology' in Patrick Lin, Keith Abney and Ryan Jenkins (eds) (n 90) 69
[112] Vladeck (n 110) 141.

with the relevant law and regulation.[113] Extracting compensation from novel beings would require them having resources that can be redistributed and effective punishment will depend on us being able to identify their interests and develop means of frustrating those interests.[114] Given our current state of knowledge about what novel beings will be like, it is unclear how the principle of responsibility should be implemented to govern novel beings with moral status if and when they are created.

5.3 Explicability

The third proposed principle for the regulation of novel beings is the principle of explicability. The principle of explicability holds that people who are significantly affected by a decision are entitled to 'a factual, direct, and clear explanation of the decision-making process'[115] used to arrive at the outcome.

One of the novel features of current techniques in AI, such as machine learning or artificial neural networks, is that the workings of the system 'are often invisible or unintelligible to all but (at best) the most expert observers'.[116] If we are going to use AI systems to make (or support) important decisions, we need to understand how these systems work. If these systems operate as non-transparent, non-interrogable 'black-boxes', then we cannot understand their decisions.[117] This is a problem because not being able to identify the causes of particular outputs will severely hamper both our ability to proactively mitigate the negative effects of using artificially intelligent systems and to ensure that these systems work safely.[118] This, in turn, could undermine trust in the system.[119]

In order for a system to satisfy the principle of explicability the explanations produced need to be intelligible to humans. It is, therefore, not enough for the system to output a log of all of the processes followed and operations performed to arrive at a particular result.[120] These results need to be presented in a way that can be made sense of by humans.[121] The principle

---

[113] Bryson, Diamantis and Grant (n 15) 288.
[114] White and Baum (n 111) 71.
[115] Luciano Floridi et al, 'AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 Minds Mach. 689, 702; MHRA, 'Good Machine Learning Practice for Medical Device Development: Guiding Principles' (MHRA 2021) principle 9. Available at: https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles accessed 12th January 2022.
[116] The Royal Society, 'Explainable AI: The Basics' (The Royal Society 2019) 8. https://royalsociety.org/topics-policy/projects/explainable-ai/> Accessed 14 July 2021.
[117] Ibid. 4.
[118] Ibid. 10.
[119] Tim Miller, 'Explanation in Artificial Intelligence: Insights from Social Science' (2019) 267 Artif Intell 1.
[120] Ibid. 11.
[121] Luciano Floridi et al, 'How to Design AI for Social Good: Seven Essential Factors' (2020) 26 Sci Eng Ethics 1771, 1781.

of explicability thus mirrors standard requirements of public reason, which also hold that people are entitled to explanations they can understand based on reasons they can accept when decisions made by others significantly affect their interests.[122] To ensure that AI systems can meet this requirement, developers of AI need knowledge of what sorts of explanations humans want and what humans generally consider to be a good explanation.

Like the principle of domination, the principle of explicability can be applied both to the regulation of precursor technologies and to the regulation of novel beings with moral status. The difference between the scenarios is not *what* is owed in each case, but *who* holds the obligation and to whom it is owed. Prior to the emergence of novel beings with moral status, the human developers of AI systems have the obligation to ensure their systems can generate explanations comprehensible to humans. If, and when, novel beings with moral status are brought into existence, the obligation could plausibly be held by the novel being itself. Should novel beings with moral status emerge, we could demand an explanation for their decisions in a similar way to how we currently demand explanation from humans who make decisions. Given the uncertainty surrounding what novel beings would be like, precisely how this could be achieved in practice is not a question we can answer prior to their emergence.

Here, as before, there are likely challenges to implementing the principle of explicability. In the case of precursor technologies, the first challenge to implementing the principle is that it could conflict with the commercial interests of developers.[123] If the algorithm being designed and the data used to train it are proprietary, applying the principle of explicability could lead to the disclosure of commercially sensitive information. A second significant challenge with its implementation is the fact that ensuring explicability is significantly harder for complex systems than simple ones.[124] As more complex systems can often perform better, the demands of the principle of explicability can conflict with other valuable characteristics of an AI system, most notably accuracy.[125] Providing an account of how to trade off explicability against accuracy is a significant challenge here.

---

[122] Onora O'Neill, 'The Public Use of Reason' (1986) 14 Political Theory 529; Onora O'Neill, *Autonomy and Trust in Bioethics* (CUP 2002) 91; Quong, Jonathan. 'On the Idea of Public Reason' in J Mandle and D Reidy (eds), *The Blackwell Companion to Rawls* (Wiley-Blackwell 2014) 268; Onora O'Neill, *Constructing Authorities* (CUP 2015) 67.
[123] Harry Surden, 'Ethics of AI in Law: Basic Questions' in Dubber, Pasquale and Das (eds) (n 20) 731.
[124] IEEE (n 90) 27.
[125] Select Committee on Artificial Intelligence, *AI in the UK: ready, willing and able?* (HL 2017-19, 100) para 99; The Royal Society (n 116) 21; Bryson (n 107) 8; Independent High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (European Commission 2019) 18. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 14th July 2021.

Harder challenges surface when we consider how to implement the principle of explicability in a post-emergence scenario. The problem is that, given that novel beings might be embodied differently to us, they are likely to perceive the world differently to us. This might make it hard for them to determine what counts as an appropriate explanation. As mentioned above, crafting appropriate explanations for decisions requires an awareness of what matters to humans.[126] The problem is, identifying what is salient to humans is notoriously difficult for artificial intelligent systems.[127] One potential solution to this problem would be to ask stakeholders what they would consider to be an appropriate explanation so that we can tell the AI what matters to us.[128] This, however, may not solve the problem if the resulting account of what matters cannot be formalised clearly enough for it to be communicated to an AI. This might be the case if, for example, what constitutes an appropriate explanation is subtly context, socially, or culturally dependent. It is, therefore, unclear whether novel beings which are embodied very differently will be able to satisfy the requirement of explicability.

5.4 Non-harm

The fourth proposed principle is the principle of non-harm, which holds that the development and deployment of both precursor technologies and novel beings should not cause harm to others. The fact that an action would cause harm to others is generally considered to be a good reason for regulating or prohibiting the activity in question.[129] The harm principle proposed here is an application of this more general moral and legal principle to the regulation of new and emerging technologies.

In a pre-emergence scenario, human developers would hold the duty to ensure the systems they design and deploy do not cause harm to others. In other words, they need to ensure their systems perform the tasks they are intended to perform, in the expected fashion, in a consistent way, without causing risks to the health and safety of humans. Given that the world is unpredictable, to ensure that systems are safe, reliable, and robust designers will need to ensure that they can manage unfamiliar events and unexpected scenarios in such a way as to avoid harms.[130] To do this, developers will need to train their models on data that is sufficiently broad,[131] stress-test their systems to see how they operate in unfavourable situations (e.g. by

[126] Leslie (n 92) 35.
[127] Miller (n 119) 12.
[128] Coeckelbergh (n 91) 2064.
[129] Joel Feinberg, *Harm to Others* (OUP 1984) 26; John Stuart Mill, *On Liberty* (Yale University Press 2003) 80.
[130] Commission, 'On Artificial Intelligence – A European approach to excellence and Trust (White Paper)' COM (2020) 65 Final 20; Floridi et al (n 121) 1776.
[131] COM (2020) 65 Final 19; MHRA (n 115) principles 3 and 8

using adversarial learning techniques or using simulations),[132] monitor their performance,[133] and introduce 'fail-safe' measures.[134]

Implementing these requirements to govern the development of precursor technologies presents some challenges. Ensuring that systems are rigorously tested before release will likely increase the costs of development; as will requirements to monitor and adjust the system once it has been deployed. Enforcing these requirements will also present challenges for regulators, who will need technical expertise to ascertain whether safety systems are up to task.

As with the other principles, things are more complicated when it comes to applying the principle to a post-emergence scenario. If novel beings develop moral status, the duty to satisfy the harm principle could theoretically be transferred to the novel being itself. There are, however, a series of challenges to implementing the principle in this way. The first problem is that given the uncertainty regarding what novel beings would be like, it isn't clear whether a novel being's understanding of the purpose of what they are doing will match that of humans. This problem is analogous to the main problem implementing the principle of explicability: i.e. it is unclear whether novel beings will share our understanding of what matters. Without an understanding of the purpose humans are trying to achieve, the novel being might be unable to recognise situations in which the sub-goals they are pursuing are no longer means to the valued end.[135] As a consequence, the deployment of novel beings might lead to adverse events and unintended perverse outcomes.

The second problem concerns enforcement. As mentioned above, it is unclear how we could enforce the duty to uphold the harm principle against novel beings. The worry here is analogous to the challenges to implementing the principle of responsibility: it is unclear how we could incentivise a novel being to follow the principle of non-harm, or how we would punish or extract remedies from a novel being should the principle be violated.[136] Until these uncertainties are resolved, it is unclear how the principle could be implemented to govern the behaviour of novel beings.

6. Concluding Thoughts: Principles, Uncertainty, and Normative Guidance

Our starting point in this chapter was to note that if novel beings come into being, the law needs to be able to take account of such beings, as well as

[132] High-Level Expert Group on Artificial Intelligence (n 75) 27; Floridi et al (n 121) 1777.
[133] Leslie (n 92) 32; Independent High Level Expert Group on Artificial Intelligence (n 125) 20; MHRA (n 115) principle 10.
[134] Robert Challen et al, 'Artificial Intelligence, Bias, and Clinical Safety' (2019) 28 BMJ Qual Saf 233.
[135] Leslie (n 92) 33.
[136] Yampolskiy and Fox (n 86) 219.

their precursor technologies. We suggested that the principal difficulty in trying to think about the regulation of (the emergence of) such beings is the high degree of epistemic uncertainty surrounding this possibility. We noted that there are (at least) four aspects to this: uncertainty regarding (1) whether novel beings will actually come to exist, (2) how this might happen if they do, (3) what the impacts of bringing such beings into existence might be, and (4) what such novel beings will be like (corporeally, cognitively, morally, and so on). Given this, we saw that one regulatory strategy is simply to wait and see. Although often favoured by regulators when it comes to uncertain technological developments, we made the case that this approach tends to lead to unsatisfactory reactive regulation later on. One consequence of this is that by waiting, the ability of regulators to influence the trajectory of technological development and deployment is significantly reduced. Equally, we noted that having a detailed regulatory regime early on would also be unsatisfactory. First, and most significantly, we simply do not have enough information to produce comprehensive and granular law and regulation to govern (the emergence of) novel beings. Second, given this, any *sui generis* regime drafted is likely to be out of date and obsolete pretty quickly. As such, what is to be done? Our suggestion in this chapter is that under conditions of significant uncertainty, a principles-based approach could offer a way to do something *now* and to offer appropriate normative guidance for the development of more concrete law and regulation *later*. Such an approach could keep our regulatory options open, helping to avoid the problem of crafting a detailed pre-emptive regulatory regime (which becomes obsolete or inapplicable as we gain more knowledge about novel beings and what they are like).

As noted by Black and colleagues, and we saw in section four, "[t]he potential benefits claimed of using principles are that they provide flexibility, are more likely to produce behaviour which fulfils the regulatory objectives, and are easier to comply with."[137] However, as we also noted, principles are not a panacea. Significantly, a principles-based approach offers us high level normative guidance and the flexibility needed to proactively regulate precursor technologies without hampering our ability to respond to new developments as they surface. Yet, as we saw in the previous section, implementing the principles of non-domination, responsibility, explicability, and non-harm, especially in a post-emergence scenario, is mired in uncertainty. Questions that would need to be answered to implement the principles outlined to govern the behaviour of novel beings include: what activities constitute arbitrary interference with a novel being? How do we ensure novel beings do not pose an unacceptable risk of harm to others? How could we hold novel beings responsible for any harms caused? How could they pay compensation? What would punishment of a novel being look like? And how could we ensure novel beings provide explanations for their decisions which are comprehensible to humans?

[137] Black, Hopper, and Band (n 40) 193.

As well as these specific problems with each principle, there is also the more general problem of how to translate the principles into practice. Given the generality at which they are expressed, principles can be ambiguous and need to be interpreted in order for them to apply to the particular situations that confront regulators. As is usually the case in principles-based regulation, different organisations will interpret the different principles in different ways (usually to their own benefit or perceived benefit). Although this flexibility is necessary in light of the epistemic uncertainties we are facing, flexibility has the downside that we can end up with actions and outcomes which are undesirable or unsatisfactory and which were not anticipated at the time the principles were drafted.

Above we suggested four potential ways in which these problems could be mitigated. The first was to supplement each principle with a fuller explanation of their purpose to help future regulators interpret them appropriately. The second was to provide an account of the principles as a collective, helping reduce ambiguity over the normative direction of travel intended. The third was to combine general principles with more specific prescriptive rules that establish hard lines that should not be crossed in applying the principle. The fourth was to complement principles with a set of best practice examples. All four of these steps to mitigating the potential vagueness of a principles-based approach rely on making principles more specific, thereby constraining the way they can be interpreted by future regulators and regulatees.

The problem, however, is that mitigating concerns about indeterminacy in the application of principles necessarily involves reducing the flexibility of the regulatory regime. Meanwhile, the benefits of flexibility are connected to the problems of interpretation and application caused by the generality of the principles. We cannot have the benefits without the burdens. The best we can expect is a trade-off between our ability to react to novel developments with a flexible regulatory regime against the risk that the lack of specificity of the principles will lead to them being interpreted inappropriately. Our suggestion is that a principles-based approach to the regulation of (the emergence of) novel beings could strike a more plausible balance between flexibility and specificity than either adopting a 'wait and see' approach (which hampers our ability to influence the development of technologies) or the creation of a *sui generis* regulatory regime (which runs the risk of becoming obsolete as technology develops).

If carefully crafted, a principles-based approach has the potential to be both flexible enough to cope with future uncertainty and normative enough to be action-guiding. The four principles suggested in this chapter (non-domination, responsibility, explicability, and non-harm) are just that: suggestions. There may be others which are better suited, given the conditions of uncertainty under which we are operating. There are almost certainly others, which we have not had the space or time to consider here, and which could be added to round out our list. What is certain, however, is that no matter which principles are used, challenges will remain. We will

need to think about: how to decide conflicts between principles;[138] who/which organisation is best suited to take on the role of regulator when it comes to emerging technologies and novel beings; whether a single regulator is even suitable given the broad range of precursor technologies and potential novel beings possible; if several regulators/regulatory agencies are to be involved, how do we ensure there is not a diffusion of responsibility and a lack of coordination between agencies. We do not propose any answers to these questions here. What is needed to begin to answer these questions is likely some sort of in-depth 'legal foresighting' process such as that proposed by Laurie and colleagues for dealing with uncertainties regarding new technologies.[139] By this they "mean the identification and exploration of possible and desirable future legal and quasi-legal developments aimed at achieving valued social and technological ends".[140] We end, therefore, simply by noting that even if we can agree that a principles-based approach could help, and on which principles are entailed, this would only be the beginning of the process when it comes to the regulation and governance of (the emergence of) novel beings.

[138] Our thanks to the anonymous reviewer for prompting us to include this concern.
[139] Laurie et al (n 33).
[140] Ibid. 3. The authors propose an in-depth framework for doing this at 11-27 and look at its application in practice at 27-32.