Disruptive Agents and Our Online World: Should We Be Concerned?

Roger Brownsword*

Abstract

Responding to Mireille Hildebrandt’s recent book, *Smart Technologies and the End(s) of Law* (2015), this essay considers whether we should be concerned about the development of “smart” agent-like machines—that is, machines having the capacity to interact with, to profile, to pre-empt, to classify, and to risk-assess human agents. Moreover, should we be concerned that these machines, in conjunction with other technologies, might replace legal rules and challenge the values of the Rule of Law? Arguing that we should be so concerned, the essential terms and conditions of any defensible social license for these machines are sketched and a new agenda for jurisprudence is introduced.

I. Introduction

Mireille Hildebrandt’s book1 is prompted by an anticipatory concern about the disruptive effects of a new generation of smart technologies—technologies that, among other things, have the capacity to profile, classify, and risk-assess human agents as well as to pre-empt human action and risk-manage human activities. The smartness of these technologies, which is largely a function of their machine learning capabilities in conjunction with their connectedness, is amply illustrated by the recent spectacular success of AlphaGo in defeating Lee Sedol, the South Korean world champion, in a five-game series of Go.

If AlphaGo, or IBM’s Watson, were simply super games-playing machines, this might be tough on human games-players but it would not raise more general concerns. However, it is clear that smart machines have applications in many domains beyond that of games, that they learn at speeds that surprise even their human developers, and that they operate in ways that are not always fully understood. Given such smart technologies, we might wonder whether there is a risk that humans will become redundant. As Yuval Noah Harari remarks, in a world of intelligent Watsons, “there is not much need for Sherlocks.”2 While Hildebrandt does not explicitly anticipate such a dystopian future, she is concerned that such machines presage a transformative (technological) agency that will be deeply disruptive in relation to law and society. In particular, Hildebrandt highlights concerns about the privacy, identity, and autonomy of human agents as they interact with smart machines, as well as concerns about unfair discrimination and about the way that

---

* Professor of Law, King’s College London and Bournemouth University.

1 Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (2015).

such machines might disrupt regulatory environments that are guided by commitment to the ideals of legality and the Rule of Law.

So, should we be concerned? Is AlphaGo the thin end of a wedge that will lead to the compromising of the conditions for human agency, and do smart technologies signal that time is up for Law? Responding to these questions, this essay presents some reflections on the use of personal digital assistants and human autonomy (Sections II and III); on the threats to both autonomy and privacy in environments where humans are exposed to smart machines that are constantly processing data about them (Section IV); on the possibility of algorithmic bias, discrimination, error, opacity, and the like, where smart machines are used to risk-assess, classify, and make decisions about the freedom or entitlement of particular humans (Section V); on the regulatory effects of smart machines, their relationship with the Rule of Law, and their possible compromising of human dignity (Sections VI and VII); and, finally, on who should be concerned about these matters.

II. Diana in the Onlife World: Should We Be Concerned (Take #1)?

Hildebrandt introduces readers to the world of smart machines, to “the onlife world,” through the eyes of a human agent, “Diana,” whose life is organized by a personal digital assistant (PDA) “that is distributed between [Diana’s] smart phone, the system running her smart house, the smart car, her ubiquitous computing office platform, while being on speaking terms with other systems, like those for traffic control and healthcare, commercial and governmental service providers, as well as monitoring systems for private and public safety and security” (1). Diana has joint custody (with her ex-husband) over her six-year-old daughter; we also know that Diana has a 96-year-old grandfather, Jacob, who is cared for by a PDA that “runs his household and monitors his health, but also serves as a social companion” (6). Despite the best efforts of these PDAs, things do not always run entirely smoothly—for example, having a PDA mediating close human relationships (such as that between Diana and her daughter) is not always straightforward; and, when Jacob suffers a mild stroke, the PDA that cares for him decides against medical examination (because, so the PDA “reasons,” this would be Jacob’s preference) only for the old man to die a couple of days later of a further stroke. Should we be concerned?

Even if predictive (automated) shopping and the like are not for us, we might feel quite sanguine about Diana’s onlife world. After all, if we were to substitute a reasonably competent flesh and blood PA for the PDA in Diana’s world, it seems unlikely that, in those situations that challenge the latter’s capabilities, the human would do any better. We have no reason to suppose that Diana’s problems would be mitigated if only a human PA were in place—or, indeed, if Diana ran her life without a PA. Similarly, it is far from clear that an empathetic human PA, rather than a smart PDA, would have averted Jacob’s death. Moreover, it seems that having a PDA is a matter that is within Diana’s control. If she prefers to live without a PDA, she can do so. The use of the PDA seems to be Diana’s autonomous choice; insofar as there is any threat to her privacy, it is one to which
she has freely consented. Accordingly, unless Diana has less choice about having a PDA than the facts suggest, or unless there is more to smart machines than meets the eye, we do not yet have a major cause for concern.

III. Diana in the Real World: Should We Be Concerned (Take #2)?
Diana has christened her PDA “Toma”; apparently, such christening is “a common habit amongst those who have entrusted a variety of tasks to their PDAs.”3 While some might worry about the relationships formed between humans and their devices,4 for present purposes, what is more material is that Diana and her grandfather are not the only PDA users in the world. Indeed, we can assume that, in Diana’s world, pretty much everyone has a PDA. If so, then Diana’s choice in using a PDA might be seriously constrained by the expectations of those with whom she works and socializes. Even if regulators make every effort to respect and protect the preferences of those who do not wish to use a PDA, where use is the norm, the choice that actually faces Diana is likely to be between participating in the mainstream onlife world or living (like the Amish) “off grid.” No doubt, the extent to which the choices of agents are constrained will vary from one setting to another and it would take some empirical analysis to establish the actuality. However, it would be dangerous to assume that there is no cause for concern simply because humans can always decline to have a PDA or otherwise engage with smart machines.

That said, even if Diana’s choice is more apparent than real, should we be concerned that her life is managed by Toma? Arguably, we should not: Toma is one step ahead of Diana in knowing what her daily preferences (for music, for food, and so on) are; and Toma’s decisions are constantly guided by what is judged to be in Diana’s best interest. On Toma’s watch, Diana will come to no harm. Nevertheless, if Toma makes a succession of substituted judgments, albeit guided by a remarkable understanding of Diana’s standard preferences, is there a risk that this might stray into paternalism? If we couple this risk with the likelihood that Diana has no real choice in using a PDA, then there is an obvious concern: is the PDA-managed bubble in which Diana lives compatible with the development of, and respect for, Diana’s autonomy? No doubt, the answer to this question will be shaped by one’s particular conception of autonomy. Given that autonomy presupposes some degree of choice and control, both as to the particular occurrent decisions one makes and as to one’s more general and enduring defaults, it follows that there must be a concern unless Toma is programmed to leave Diana to make the choice when it really matters (cf. 65-68, where a PDA diagnoses a health risk and makes an intervention of some kind). To the extent that PDA technology affords Diana the option of reserving to herself certain matters for decision, this eases some of the concerns; but if we want to know which choices should “really” matter to agents, we enter

3 Hildebrandt, supra note 1, at 2.
4 Cf. Sherry Turkle, Alone Together: Why We Expect More from Technology and Less from Each Other (2011).
the difficult conceptual terrain of authenticity. Moreover, if we equate autonomy with moral self-development and agency, we need to take a hard look at who is making the choices (see also Section VII).

**IV. Beyond the Bubble: Should We Be Concerned (Take #3)?**

Toma can shield Diana against the activities of at least some other smart machines. However, if Diana sometimes acts outside Toma’s protective bubble—in, so to speak, “Big Data Space” (46)—should we be concerned? To the extent that she so acts, Diana seems to retrieve some autonomy. But at what cost? For example, is there now a threat to Diana’s privacy? Likewise, are there threats beyond the bubble for humans who, unlike Diana, do not enjoy (cannot afford) an adequate level of protection?

Hildebrandt defines privacy as a right to be free from “unreasonable constraints on the construction of one’s own identity” (80). With privacy so defined, by breaking free from Toma’s protective bubble, Diana is not only asserting her right to autonomy but also her right to privacy; when Diana returns to her bubble, Toma might need to reassess who Diana is. Yet, outside the bubble, Diana cannot be assured that she will be free to play around with her identity; both inside and outside the bubble, Diana might encounter unreasonable constraints that infringe her privacy. Where smart machines interact with human agents, gathering data about them, mining the data, and then initiating some “response”—whether a precautionary alert or a nudge or even a degree of incapacitation, and whether initiated for the benefit of the respondent or in the interests of others—some constraints will seem more reasonable than others.

While there might be some opportunity for Diana (or for Toma, on behalf of Diana) to negotiate the terms and conditions of her interactions with smart machines, we can assume that public regulators will put in place some basic ground rules for the collection and use of data. For example, there might be rules that follow European data protection principles by aiming to make the collection and processing of data transparent and by confining use to purposes that meet some test of legitimacy and/or consent (cf. ch. 9). So much for the outward flow of data but, outside her protective bubble, Diana might find that she is the recipient of information that she would prefer not to have. Suppose, for example, that Diana (who, we know, enjoys a glass of wine) is attempting to limit her intake of alcohol by having some “dry days.” If, on her dry days, Diana finds her resolve weakened by a succession of special offers, delivered by machines that are smart enough to know when Diana is most vulnerable and likely to succumb, then might this kind of targeting be a violation of both her autonomy and (recalling Hildebrandt’s definition) her privacy (see 80, 92-93)? In other spheres—notably where genetic information is returned to patients, to research participants, to family members, to the Havasupai people, and so on, regardless of their wishes and without respect for a claimed “right not to

---

5 Cf. Cass R. Sunstein, Choosing Not to Choose: Understanding the Value of Choice 119-20 (2015) (recognizing that there might be an argument to be made for active choosing when “learning, authenticity, responsibility, and the development of values and preferences are important”).
know”—might we also treat such acts as an infringement of an agent’s identity-related privacy? Arguably, in all worlds, offline, online, and onlife, privacy qua the right not to know might prove to be extremely important.

V. Smart Machines and the Assessment of Risk: Should We Be Concerned (Take #4)?

One of Hildebrandt’s central points is that, generally speaking, humans do not have difficulty in “reading” one another: one agent can figure out how others perceive him or her. However, in the onlife world, where human agents are interacting with smart agent-like machines, the former know that the latter are constantly building and revising profiles of them but without being able to read the latter. While, as Hildebrandt argues, this might present some difficulties for agents who want to understand how they look through the eyes of fellow agents, this is surely not the major cause for concern—or, at any rate, not for mature agents. Rather, what will concern human agents is the possibility that smart machines will make decisions that will treat them inappropriately as a “bad risk” (e.g., for credit or employment) or as “high risk” (e.g., for criminality or insurance). Although smart machines will constantly work to reduce their rate of errors, mistakes will be made. Rightly, this will be a cause for concern and particularly if the errors that are corrected tend to be the false negatives rather than the false positives. For example, as Andrea Roth has forcefully argued, where there is strong pressure for effective crime control and, concomitantly, a tendency for politicians and criminal justice professionals to be more concerned about false negatives (about the guilty who escape prosecution, conviction, or punishment) than false positives, we can expect there to be “a desire for a particular type of accuracy: the reduction of false negatives.”

First, though, it is important to understand that there is likely to be a recurrent tension between the “reason” or “intelligence” that drives the machine and the demand that human agents make for individual justice. Suppose, for example, that Diana, wishing to upgrade her smart car, applies for a credit facility but that she is turned down by a smart machine that classifies her as a bad risk. When Diana challenges the decision, she learns that one of the previous occupiers of her house had a record of non-payment of loans. But, why, Diana asks, should the credit record of an unrelated third-party count against my application? Is that not unfair and irrational? To which the response is that the machine makes more accurate decisions (i.e., decisions involving fewer false negatives) when it uses third-party data in this way and that, if such data were to be excluded from the calculation, the cost of credit would increase. In fact, this is not a novel issue. In the English case of *CCN Systems Ltd v. Data Protection Registrar*, on facts of this kind, the tribunal held that, while it accepted that such third-party information might have general predictive value and utility, its use was unfair to the individual and could not be permitted.

---


Similarly, Diana might argue that she has been treated unfairly if her application for credit is successful but the terms and conditions of the facility reflect the fact that (because of unrelated third-party data) she is classified as a higher-than-average risk and, once again, the response will be that the costs of credit will be increased if such data is excluded. How is the choice to be made between the general utility of the credit algorithms and the unfairness of particular decisions?

Before answering that question, we should note that, in our hypothetical, Diana at least knows that a decision has been made and, when she challenges the decision, she is given a reason. Things could be much less satisfactory: for example, Diana might not be told that her application has been rejected (she simply gets no response or some misleading reply), or she is given no reason for the decision, or she is told (quite honestly) that it is not possible to explain how the “black box” operates (although it is known that, within the black box, there are processes that constantly work to improve the performance of the system). As Frank Pasquale has remarked, “The black boxes of finance [have] replaced familiar old problems with a triple whammy of technical complexity, real secrecy, and trade secret laws.” Even if Diana had the opportunity to articulate her grievance, even to present it to a human for final determination, she would find that the efficiency and utility of the algorithms is pitted against her pleas for transparency and due process.

With the question (concerning utility or fairness) still pending, there is one more point to make. It is one thing for a smart machine to deny Diana access to credit, but smart machines might make exclusionary or pre-emptive decisions that have far more serious consequences for Diana. For example, a smart machine might decide that Diana’s daughter is “at risk” and needs to be taken into care, with both Diana and her ex-husband being denied access to their child. Or Diana might be denied access to the light rail that she seeks to take to the office because she is judged to be a security risk. Although (in this book) Hildebrandt does not discuss the operation of the criminal justice system in a sustained way, it is obvious that this is one area where smart machines might be deployed—for example, to enable policing to be more “intelligent” and the use of resources more efficient but also to take pre-emptive action against agents who are judged to be high risk, to deny bail to arrestees who are assessed as high risk, and to extend custodial terms for offenders who, at the point of release, are still judged to be “dangerous.” In all these cases, smart machines churn out decisions that are in line with Benthamite principles and that are generated by the logic of big data but that can carry through off-line discriminatory practices and depart from the ideal of a “justice” system.  


9 See also Mireille Hildebrandt, Proactive Forensic Profiling: Proactive Criminalization?, in The Boundaries of the Criminal Law 113 (R.A. Duff et al. eds., 2010).

10 For a seminal three-pronged critique of such an “actuarial” approach to criminal justice, see Bernard E. Harcourt, Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age (2007). While Harcourt
are questions being raised in the US about the hidden racial bias of apparently colour-blind algorithms used for bail and sentencing decisions.\textsuperscript{11}

In response to the question that we keep confronting, the choice is one for each society to make. As Hildebrandt remarks, we need to ask the ultimate political question of “what kind of society we wish to remain or become” (172).\textsuperscript{12} Accordingly, smart machines should be regulated by whatever terms and conditions are specified in the particular social license that permits their use. How are such terms and conditions to be specified? Within a constitutional democracy of the kind that Hildebrandt takes to be the benchmark, there will be deliberative processes that lead to some “acceptable” balance of the competing interests (in utility and in fairness). Wholly “unreasonable” views will be excluded but this is not to say that everyone will judge the particular agreed-on outcome to be the most reasonable amongst the options that are acceptable. While this serves to democratize the ground rules for the use of smart machines, it also seems to relativize them. Are there no necessary limits in each society’s license?

If we assume a community of human agents with moral aspirations, we immediately have a triple bottom-line for regulators. The terms and conditions of any social license for new technologies should be such as to preserve, protect, and promote:

- the essential conditions for human existence (given human biological needs);
- the generic conditions for human self-development and agency; and
- the essential conditions for the development and practice of moral agency.

The point about these three bottom lines is that the “conditions” (whatever they are agreed to be) are regarded as neutral as between one human and another, between one agent and another, and between one agent with moral aspirations and another. These conditions represent a “commons” that reflects the needs of all humans, irrespective of their particular projects and plans as agents, and irrespective of their particular moral beliefs. Accordingly, regulators should reject as “wholly unreasonable” any proposal for the use of smart machines that might compromise any of these bottom-line conditions. Of course, determining the nature of these conditions will not be a mechanical process and I do not assume that it will be without its points of controversy. Moreover, even if a community agrees upon where the bottom lines are to be drawn, it would still have to decide how to handle proposals for above-the-line uses of smart machines (such as in the example of Diana’s application for credit), which is precisely where the choice between general utility and individual justice has to be made.


\textsuperscript{12} Echoing Pasquale, supra note 8, at 216.
Significantly, it is in relation to such contested choices that Law enters the picture, playing a distinctive role in facilitating the debate and determination of moral differences. In a community with moral aspirations, the social license for smart machines will be translated into a legal framework. That framework might be guided by more than one moral approach, sometimes reflecting a somewhat pragmatic compromise between competing views, and sometimes a long way short of ideal. Accordingly, in such a scenario, when the Law commands respect for its rules, it is saying more than that it would be prudent to comply but less than that the legal position comes warranted as the expression of moral truth. Following this line of thinking, what the Law represents is the community’s good faith and best attempt to settle its moral differences. In principle, disagreements about the legitimate applications of smart machines should be settled in this way; but, it will be recalled, one of Hildebrandt’s most profound concerns is that smart machines will emerge as regulatory rivals and threaten to displace the Law.

VI. Is This the End of Law? Should We Be Concerned (Take #5)?

Although Hildebrandt shares the common convention that when we speak about the Law, we refer to “an institutional normative order” (143), she has a very particular conception of Law. Amongst the distinctive characteristics of Hildebrandt’s conception of Law are: that legal standards are co-produced (reflecting a commitment to participatory and inclusive democratic practices); that the “mode of existence” of modern Law, with printing technology providing its infrastructure, is in the form of texts (statutes, codes, precedents, and so on); that legal texts are open to interpretation and contestation (in courts) before their application in individual cases; and that these features, in combination, enable Law to respond to more than the demand for certainty or, in Hildebrandt’s terminology, mere “legalism.” In the ideal-typical case, Law is also able to instantiate “legality” by responding to the demand for individual justice and for legitimate purposes (see, e.g., 154-55). By contrast, where order is controlled by technological regulation, we find a very different story. First, technological regulation is not “controlled by the democratic legislator and there is no legal ‘enactment’”; second, the design of technological devices might be such as to “rule out violating the rule they embody, even if this embodiment is a side-effect not deliberately inscribed”; and third, “contestation of the technological defaults that regulate our lives may be impossible because they are often invisible and because most of the time there is no jurisdiction and no court” (12, with the use of cookies as an illustrative example; see also 183-85). Stated shortly, Hildebrandt’s concern is that smart machines will enhance the power and expand the possibilities for technological regulation in a way that crowds out the features that we value in Law.

To some extent, the use of smart machines as regulatory tools simply exacerbates long-standing concerns about the capacity of powerful private interests to escape and evade public accountability and oversight. The fear is that transnational corporations, 13 For this version of “legal idealism,” see Deryck Beyleveld & Roger Brownsword, Law as a Moral Judgment (1986).
playing by their own rules, will be able to draw on smart machines to impose further regulatory effects. However, as Hildebrandt recognizes (see ch. 10), legal rules alone might not be sufficient in constructing an acceptable regulatory environment for transactions and interactions in the onlife world; we might also need to call on “Legal Protection by Design.” What should we make of this?

First, for a community with moral aspirations, the reciprocity that is implicit in Hildebrandt’s analysis of the Rule of Law (together with the ideal of legality) is of fundamental importance. On the one side, those who make the laws should do so in ways that are inclusive, transparent, and faithful to the community’s moral commitments and, on the other, those who are subject to the laws should respond to the laws as products of a process that merits respect. Law, on this view, is always a work in progress; but its call for respect is necessarily moral (in the sense indicated above). Second, all the standard versions of the Rule of Law presuppose that Law is expressed in the form of rules. Although the familiar laundry lists for legality can be given some application to technological regulation, we need to focus, as Hildebrandt does, on the framework for law-making. Third, when a proposal for the use of technological regulation (or Legal Protection by Design) is brought to a legislative assembly, there needs to be debate, as usual, about the purpose to be served. However, the distinctive questions to be debated are whether use is compatible with the triple bottom-line for regulators and, in particular, whether eliminating the practical option of non-compliance is compatible with the community’s moral aspirations. Even if smart machines signal the end of a good deal of Law as we know it—with chunks of “regulatory” criminal law, contract law, and tort law becoming redundant as technologies directly manage matters of health and safety, environmental protection, and so on—it does not follow that we must abandon the values of the Rule of Law. Provided that technological regulation is subjected to a reworked understanding of these values, smart machines should make a legitimate, licensed, and lawful contribution to the ordering of the onlife world.14

VII. What About Human Dignity? Should We Be Concerned (Take #6)?

When new technologies emerge, they are often greeted with concern about their impact on human dignity. Generally, commentaries on developments in ICTs have not highlighted such a concern, but the development of intelligent machines might well provoke a new concern about the compromising of human dignity.15 To some extent, this might be a position taken in above-the-line debates but it is important to understand the place of human dignity in bottom-line debates about the conditions for any aspirant moral community.


15 See Roger Brownsword, From Erewhon to AlphaGo: For the Sake of Human Dignity, Should We Destroy the Machines? (Keynote, Conference on “The Future of Human Dignity,” Utrecht University, 2016).
For moralists, it is axiomatic that each agent should develop a sense of what it is to do the right thing for the right reason, and try always to do just that. Human dignity involves more than merely acting in line with the right thing; the paradigmatic expression of the dignity of humans is in doing what an agent judges to be the right thing even where there is an opportunity to do the wrong thing. In the light of this, it is one thing for agents to use their PDA as a moral “critical friend,” quite another to delegate moral decision-making to their PDA (or to habitually rely on the PDA’s “expert” judgment); it is one thing for an agent freely to comply with legal rules but quite another to have no practical option other than to comply with the constraints imposed by whatever technological measures have been adopted. Where agents no longer freely make and act on their moral judgments, we should question whether the conditions for moral community—and, concomitantly, for human dignity—are being compromised.

Yet, is it plausible to think that a particular employment of technological management will make any significant difference to the context that is presupposed by moral community? If the fitting of locks on doors, or the installing of safes, and the like, has not fatally compromised the conditions for moral community, why should technological management aided by machine learning be any different? Granted, today’s technologies are varied, sophisticated, and dense; but, is this not simply a difference of degree? Surely, there still will be sufficient occasions left over for agents freely to do the right thing and to do it for the right reason as well as to oppose regulation that offends their conscience. On the other hand, turning up the temperature from cool, to warm, to hot, is also just a matter of degree; but it does not follow that we are comfortable at all points on the scale. Accordingly, it will be for each community with moral aspirations to make its own assessment of the conditions that are required for it to flourish and, where there is uncertainty about this matter, it will need to judge how precautionary it should be in licensing the use of technological regulation.

VIII. Finally, Who Should Be Concerned?

Finally, who are the “we” who should be concerned? The short answer, surely, is “everyone.” Those who are developing smart machines need to conduct their research in a responsible way. There need to be sustained attempts to engage the public. There might need to be commissions to coordinate debate. And, politicians and regulators alike need to be mindful of their responsibilities. However, the “we” that I particularly have in mind is the community of jurists.

It is jurists who need to concern themselves with the use of technological management (including by smart machines) that seems set to join law, morals, and religion as one of the principal instruments of social ordering and control.16 To a considerable extent, technological infrastructures that support our various transactions and interactions will structure social order. The domain of legal rules is set to shrink. The implications for a

---

jurisprudence that is predicated on the use of rules and standards as regulatory tools or instruments should not be underestimated.

If we are to bring the issues raised by technological management onto the jurisprudential agenda, it is not enough to broaden the scope of (Westphalian) Law in the way that transnationalists and pluralists propose; Hildebrandt is right to resist a purely instrumental focus on “regulation.” Rather, we should start with a concept of the regulatory environment that accommodates both rules and technological management. With the field so drawn, we can re-energize thinking, just as Hildebrandt does, about the Rule of Law and the virtue of legality. We can take stock of the changing complexion of the regulatory environment. And we can assess the significance of these technological developments for traditional legal values as well as for the communities of human agents who live through these transformative times.

---
