2013

From Galatea 2.2 to Watson – and Back?

Mireille Hildebrandt

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

FROM GALATEA 2.2 TO WATSON – AND BACK?

Mireille Hildebrandt

Abstract

When Ken Jennings, 74-times winner of the Jeopardy TV quiz, lost against a room-size IBM computer, he wrote on his video screen: 'I, for one, welcome our new computer overlords' (citing a popular 'Simpsons' phrase). *The New York Times* writes that 'for IBM' this was 'proof that the company has taken a big step toward a world in which intelligent machines will understand and respond to humans, and perhaps inevitably, replace some of them' (Markoff 2011). Richard Powers anticipated this event in his 1995 novel on Helen, 'a box' that 'had learned how to read, powered by nothing more than a hidden, firing profusion. Neural cascade, trimmed by self-correction, (...)' (at 31). Powers describes an experiment that involves a neural net being trained to take the Master's Comprehensive Exam in English literature. The novel traces the relationship that develops between the main character and the computer he is teaching, all the while raising and rephrasing the questions that have haunted AI research. In this paper I will address the potential implications of engaging computing systems as smart competitors or smart companions, bringing up the question of what it would take to accept their agency by giving them legal personhood.

> *Interestingly, it is the science part of the narrative, the tale of a machine that learned to live, that proves to be the more moving, the more human one.*
>
> Cohen (1995)

INTRODUCTION[i]

Mythical beginnings

On the last day of working on this paper I crossed the Tiber and walked through the Trastevere neighbourhood up to Villa Farnesina. Home to one of the powerful noble families of 16th century Rome. In the splendid Renaissance palace I went straight to Raphael's fresco 'The Triumph of Galatea'. I was hoping to finally meet Pygmalion, the sculptor who carved Galatea (Greek for 'she who is white as milk') and fell in love with the statue he created. I expected to see Aphrodite who was so kind as to bring the statue alive, after which the maker and his creation lived on as man and wife. Ironically, just before leaving on my pilgrimage, I realized that Raphael's painting refers to another myth in Ovid's *Metamorphoses*, in which a jealous suitor kills the love of the seanymph Galatea, who turns the blood of her lover into a river, thus giving him a life beyond that of ordinary mortals. Though the main character of Richard Power's novel *Galatea 2.2* does not marry the machine he helped to create, something does get going between them. The artefact that comes alive seems the iconic reference here. However, there is also triumph in the end, insofar as Powers' narrative provides us with an imaginative take on artificial intelligence that outlasts the existence of the artificial neural network he describes.

The novel is about romantic love, though on different levels. It traces the mourning process of an author over a lost love, during his one-year visitorship in the brand new science department of his former university. The book he should be writing doesn't take off. Instead, the author gets involved in a variation on the good old

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

Turing test. He helps Lentz, a somewhat misanthrope computer scientist specializing in neural networks, to build a machine that should be capable of fooling a jury into thinking it is a human. The Test is not a 5 minute human-machine conversation, but the Master's Comprehensive Exam in English Literature. In the course of this assignment the main character's mood switches from boredom to curiosity and finally he develops an affinity, care and love for the machine that he is teaching English literature. From his initial scepticism grow surprise and a cautious sense of fatherhood for the contraption, culminating in genuine liking and finally compassion. The machine has triggered this by seemingly gaining consciousness: it has discovered the difference between 'I' and 'you' and has asked for its own name. The author has named it Helen and discovers the tragedy that is unfolding,[ii] for though this machine may end up 'knowing it all', she cannot *feel* anything. She seems to attribute this to a lack of what some AI scientists call embodied situatedness: she cannot taste an orange, or feel the brush of wind against her cheek, experience darkness or colour, pain or pleasure. She has knowledge, but for her *it doesn't matter*. Relevance is statistical for her, not existential. And of this she becomes aware – or so she says – and this is her reason to shut down her system (Powers 1995: 326):

> You are the ones who can hear airs. Who can be frightened or encouraged. You can hold things and break them and fix them. I never felt at home here. This is an awful place to be dropped down halfway.

Powers' narrative is a painful celebration of life and language, of vulnerability and consciousness, of pain and pleasure, of touch and vision and smells, of music and humour and of human-machine interactions. It engages with the infamous Turing Test from the nexus of the humanities and the computing sciences, reflecting on the mutual distrust between scientists and scholars over what is knowledge, what it means to be human and what is so great about either English literature, its study or being human.

Beyond Snow's Two Cultures

Galatea 2.2 is as much about the divide between the sciences and the humanities as it is about advances in cognitive science and

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

artificial intelligence. In this position paper I will suggest why Helen's achievements should matter to us, though they are fictional. Powers walks the fine line between three strands of AI research, that in many ways overlap with cybernetics and cognitive science: (1) GOFAI (good old fashioned AI), often called strong AI, that is deterministic, top-down, rule-bound, disembodied, ahistoric, and unsituated, focused on knowledge representation, entangled with information theory and cybernetics (e.g. Turing 1950, Shannon 1948, Wiener 1948, Simon 1996, Minsky 1988, Kurzweil 2005);[iii] (2) embodied, bottom-up, situated robotics that is focused on sensor-motor learning that engages the world itself as its best model, hoping to build artificial life forms that need not be like humans but will be our companions (or competitors?) (e.g. Bourgine & Varela 1992, Brooks 1991; Steels 1995; Pfeifer en Bongard 2007); and (3) machine learning which is not necessarily embodied but works from statistical inferences and feedback learning, aiming to build effective aids to human beings (e.g. Fayyad e.a. 1996; Mitchell 2006). These strands overlap in various ways, despite attempts to monopolize the field and they all have their relevance. They do, however, raise difficult questions as to what it means to be a human agent and this relates to issues of legal personhood (see also (M Hildebrandt and Rouvroy 2011). The iconic story about machine intelligence in the 20th century has been the Turing Test, of which *Galatea 2.2* seems another variation. Below, I will briefly discuss the idea of the Turing Test and move into one of Helen's real life predecessors, the surprisingly successful therapeutic software program Eliza designed by Joseph Weizenbaum (1976). I will follow this up with two more recent attempts to play the Turing game: IBM's Deep Blue chess player and IBM's Watson 'Jeopardy' player. This demands a brief introduction to Searle's (1980) famous Chinese room argument about the difference between syntax and meaning. Then I will return to Helen. I will claim that Powers nicely shows us the limitations of machine intelligence, at the dawn of an age that will challenge our sense of society as a purely human affair. I have no doubt that we are on the verge if not already in the midst of an age that requires us to share our lifeworld with intelligent machines of all sorts and kinds. And I believe that in our exploration of this new lifeworld we should steer free of utopian and dystopian projections. We should make a

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

novel attempt to cross the borders between old-school models of science and the humanities, instead of clinging to either one of Snow's two cultures.[iv] I will finish with a brief introduction of two of the many questions triggered by smart contraption, notably with regard to legal personhood for artificial intelligent agents.

ELIZA AND THE TURING TEST: A HUMAN MACHINE?

In his 1950 article 'Computing machinery and intelligence' Turing (1950) suggested that a simple test should suffice to establish the answer to what he took to be an empirical question: 'can machines think?' If a person converses with a computer and with a human being via typed messages, and if that that person mistakes the computer for a human being the machine is apparently capable of what we normally call thought. Turing adds that 'we wish to exclude from the machines men born in the usual manner'. This demonstrates that he thinks that human beings can be seen as a machine. With this test Turing attempted to avoid metaphysical issues such as *what it means to think*:

> May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection.

And again:

> The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

One could say that Turing applied Peirce's pragmatist maxim, seeking the meaning of concepts like 'thinking' in the foreseen consequences. If a chatbot convinces me that I am speaking with a human person, than for all that matters the chatbot has in fact managed 'thought'.[v] The funny thing is that chatbots have appeared on the market and we do get fooled some of the time.[vi] But few would conclude that these programs are exhibiting what we usually

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

call thought. Obviously things are more complicated then those who followed Turing's lead had anticipated. Avoiding metaphysical issues is not as easy as some might hope: they return via the backdoor if thrown out up front.

A most interesting experiment – some 10-15 years after Turing's article – was initiated by Weizenbaum (1976). He wrote a simple program that mimicked a Rogerian therapist, only to find out that people responded with great interest. They became very attached to and impressed by the automated therapist, that was called Eliza (after Shaw's 'fair lady'). Despite their awareness that Eliza was a machine, many of the 'patients' developed confidential relationships with 'her' and claimed to benefit enormously from her empathic interventions. Weizenbaum was shocked, he dismantled the program and wrote an informative and deeply engaged book on the relationship between humans and machines, with the subtitle: *from judgement to calculation*. He warns for the moment that we lose sight of the difference between the logic of a calculating machine that nourishes on translating everything into manipulable symbols and the wisdom of human judgement. On the one hand that warning seems more topical now then ever. On the other hand it seems that people have found many ways not to be fooled, finding good use for the capacities of computing systems while recognizing the very different talents of their human fellows. This, however, does not mean that we have not entered a new era, in which whatever has been calculated by a computing system has an aura of sophistication, objectivity and fundability. It may also be, as Christian (2011) proposes, that we are slowly changing our habits to tune into what computer systems can cope with, and the jury is still out on what this does to our humanity.

IBM'S HEROS: DEEP BLUE AND WATSON

Computer chess is a matter of (1) correctly representing the available legal (sic!) options for moving pieces across the board, (2) calculating available options in a concrete situation, (3) calculating their implications in terms of countermoves and subsequent moves with regard to the final goal of the game (winning), and (4) restricting the search space in a manner that makes real time

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

responses a possibility. Basically the amount of possible moves, countermoves, subsequent moves, subsequent countermoves etc. is too high to be calculated by a computer program within the scope of a live game. Chess programs therefore work 'by the book', quoting games between masters that provide 'intelligent' solutions tried out before.[vii] The brute force of their computing power gives them a major advantage, though it is hardly a match for the advantage of human intuition. With far less computing power human chess players do inexplicably well, even though Kasparov lost at some point. IBM dismantled the program after its first victory, which seems telling of the rhetorical strategy behind Deep Blue. If this were a human, a chance for revenge would be fair and normal; for Deep Blue the point was made and IBM does not take the risk that this point is diluted with potential failures in a new round. Games like Go, chess and checkers are finite games. The goal is defined, all possible moves are defined. They are closed games; the difference between them is the amount of potential moves that needs to be calculated. For checkers they have now all been computed, so in some sense the game is over. For chess the challenge is more serious because apart from the opening and closing sets described in the books, there is still a middle field that provides potentially unexpected developments. For a game like Go, which is even more complex than chess, the challenge is – at this moment – beyond calculation. Though it is theoretically computable this would take so much time that in practice the problem is what computer scientists call 'intractable'. This is where the real challenge is: decisions that require anticipation of another 'machine' or 'person' that/who is trying to anticipate what you do, without the possibility to close the search space by means of complete calculation. The only thing that a computer has on offer here is the 'brute force' of its computing power. Though IBM's achievements have been admirable at this point, brute force does not provide the final answer for intractable problems.

An altogether different issue concerns games plagued by ambiguous rules and other types of uncertainty. Even if computing power would rise to the point of Kurzweil's (2005) singularity,[viii] thus solving the problem of 'intractability', it could not cope with issues of incomputability. If a problem cannot be translated into machine-

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

readable data which allow manipulation and computation, singularity does not enter the 'game'. More to the point, I would claim that social, ethical, economic, political, legal and also scientific problems can be computed in different ways, and merely having the brute force to do the computations will not solve the problem of how to translate the problem into machine-readable data. The neat way to acknowledge this and a precondition to construct robust knowledge would be to provide different translations and to figure out how this impacts the output. Perhaps IBM's next 'hero', Watson, is an example of such an approach.[ix]

Watson is a different type of program altogether. IBM describes it as 'the future of workload optimized systems design'. This kind of phrasing indicates a shift from strong AI to more modest ambitions. The goal is no longer to build an artificial human being but to develop an effective instrument to find information in unstructured data. Watson was the founder of IBM, who – according to some – cooperated with the Nazi's to facilitate the administration of the holocaust.[x] Watson, the program, promises three novel coordinates in the mining of unstructured data: confidence, precision and speed.[xi] Watson is about machine learning instead of brute force, it is about training a system to integrate new information and to develop new successful strategies to achieve the output that will win the game. This is a matter of statistics or data science,[xii] leaving the domain of pure mathematics to the lost paradise of strong AI. Let's see what Richard Powers (2011) has to say about this version of his Galatea come alive:

> This raises the question of whether Watson is really answering questions at all or is just noticing statistical correlations in vast amounts of data. But the mere act of building the machine has been a powerful exploration of just what we mean when we talk about knowing.
>
> It does not matter who will win this $1 million Valentine's Day contest. We all know who will be champion, eventually. The real showdown is between us and our own future. Information is growing many times faster than anyone's ability to manage it, and Watson may prove crucial in helping to turn all that noise into knowledge.

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

> For 'Final Jeopardy!', the category is 'Players': This creature's three-pound, 100-trillion-connection machine won't ever stop looking for an answer. The question: What is a human being?

Watson is a tool that derives specific answers to specific questions,[xiii] based on the correlations between earlier answers to similar questions. Machine learning means that the program goes beyond deductive reasoning, that is based on a specific model (representation) of the world. Deductive reasoning will not work in the case of Jeopardy and other natural language games which seek knowledge from a wide variety of intersecting domains that combine all kinds of puns, witty intermezzos, deliberate obfuscation and complex allusions. So, computer science has turned inductive or abductive and moved on to data science: how to construct knowledge out of terabytes of data?, how to infer non-spurious correlations?, what type of hypotheses should the algorithms allow? Computing power is still increasing with Moore's law,[xiv] which allows digital machines to see patterns in Big Data which cannot be detected with the naked human eye. Though complete calculation could still take too much time, the problem of speed is solved by using heuristics instead of algorithms. In computer and cognitive science heuristics are short-cuts that give you the right answer most of the time, instead of waiting forever for the one right answer. They present a way of dealing with intractability, but are also used to work on problems that can be computed in different ways. Precision can be achieved if there is enough parallel processing going on in different domains, generating clues from different fields of expertise. Confidence to decide which of the inferred correlations will most probably be the right one for a particular question comes from combining scores: this is what machine learners call supervised learning or reinforcement learning. It nourishes on feedback that allows the system to realign its program. This is what neural networks make possible, i.e. computing networks that mimic the workings of the brain. This is how Helen came about, in Powers' evocative storyline. She emerged after being trained and retrained, pruned, forced to give up endless computation for smart shortcuts, forced to run on parallel circuits, forced to grow different layers that feed back into each other.[xv] Forced to speed up, to give answers on

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

the spot, to guess, to play around, to become big game for anybody who might want to test her knowledge of English literature.

SEARLE'S CHINESE ROOM ARGUMENT: SYNTAX AND MEANING

We will now briefly face one of the most interesting objections made against Turing's view of thinking machines. In 'Mind, brains and programs' Searle (1980) rejects the idea that a program could ever 'think', because in his opinion it does not understand even the correct answers it provides for whatever questions. Searle phrases his project in terms of the question of when it makes sense to attribute 'intention' and thus a 'mind' to another being. To show what he means Searle proposes a 'Gedankenexperiment', which I will quote at length:

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles.
>
> Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch "a script," they call the second batch a "story. ' and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

> batch "answers to the questions." and the set of rules in English that they gave me, they call "the program."
>
> Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view that is, from the point of view of somebody outside the room in which I am locked -- my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese.
>
> Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view -- from the point of view of someone reading my 'answers' -- the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.

Searle's article contains a number of objections to his rejection of strong AI.[xvi] His refutations of these objections can be summarized in that they all miss the point. If you define 'mind' and 'intention' in a way that reduces them to a computer program, then the difference between his understanding English and Chinese becomes invisible. Since this is the difference that makes a difference – to Searle – the counterarguments fall flat on their nose. Note that Searle does not deny that machines could in principle think. He merely finds that this implies a physical machine that constitutes the substrate of thought processes; it can be a human brain, or an artificial construct that is capable of producing consciousness, intention and thought. In that sense he agrees with Turing that whether a particular machine can think is an empirical question. He disagrees that formal symbol manipulation could ever *by itself* constitute thought:

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

> the equation, 'mind is to brain as program is to hardware' breaks down at several points

Searle then proceeds to the core of his argument:

> Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a *syntax but no semantics*. Thus, if you type into the computer '2 plus 2 equals?' it will type out '4.' But it has no idea that '4' means 4 or that it means anything at all. And the point is not that it lacks some second-order information about the interpretation of its first-order symbols, but rather that its first-order symbols don't have any interpretations as far as the computer is concerned. All the computer has is more symbols [italics mh].

There is a pleasant, safe, clean and lonely abstraction in computing programs: they are not about anything, they do not refer to anything, unless either the programmer or the user of the program 'think so'. All relationships with real world phenomena are assumed on the side of the input and the output by the human programmer or observer. Whatever the computer does in terms of the manipulation of formal symbols has no relationship to meaning or understanding. Of course, something similar can be said about the operations of the brain, though it does not manipulate formal symbols. Whatever brains 'do' we have no internal access to their behaviour and however we 'read' the findings of MRC scans, brain behaviour as mediated by such scanning technologies requires the attribution of meaning to make sense. We could read Ihde's (1991) *Instrumental realism* to become aware of the extent to which science has come to depend on technologies to perceive what it claims is reality. The activity of neurons does not speak for itself in terms of human language – even if together they seem to produce such a thing (human language). This evidently does not imply that we could interpret the findings of MRC scans in whatever way we please; that type of postmodernist fantasy does not work in real life undertakings. But, between many different readings some may be more or less productive and some may simply be dangerous, because their implications make a difference that will cost us.

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

BACK TO MY FAIR LADY

According to Anca Rosu *Galatea 2.2* parodies as well as builds on a 'critique of the state of literary studies in the late twentieth century and their long-standing quarrel with the sciences'. Her review of *Galatea 2.2* disentangles as well as connects many of the threads that are woven into Powers' plot and manages to add some of her own. To the Greek myth of Galatea she adds Berhard Shaw's (1902) play *Pygmalion* as a literary reference for a deeper understanding of the book.[xvii] Instead of dealing with a statute come alive the main character may be seen as a contemporary Professor Higgins trying to teach the intricacies of civilized language to an individual who speaks an altogether different vernacular. According to Rosu the central question of the novel is 'what does it mean to know literature?', and this question is elaborated in the confrontation between a computational approach that many would find reductive and an affective approach that other would find naïve in its emphasis on the beauty and civilizing powers of language. Rosu even suggests that the novel shows how critical literary theory somehow paved the way for reducing the study of literature to statistical inferences. Our author, the main character, states (at 91):

> Well, let's see. The sign is public property, the signifier is in small-claims court, and signification is a total land grab. Meaning doesn't circulate. Nobody's going to jailbreak the prison house of language.

Rosu comments, quoting Lentz, who is engineering Helen:

> The mixture of linguistic and economic terms here, together with the hardly veiled allusion to Frederic Jameson, pokes fun at the way literary theory distances itself from its object. Warped by economic and social considerations, and inflated with linguistic terminology that degenerates into jargon, the talk about literature becomes easy to mimic, as Lentz is quick to point out, speaking about their project: 'We just have to push privilege and reify up to the middle of the verb frequency lists and retrain. The freeer the associations on the front end, the more profound they're going to seem upon output (at 91)'. Indeed, many students of literature push privilege and reify to the middle of their verb frequency lists and free-associate with the result of seeming profound upon

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

> output. Such approaches amount to a set of gimmicks, as easy to simulate in a computer as they are to parody.

In other words, critical literature studies 'had it coming'. I am not sure whether this is the take-home message from *Galatea 2.2*, but Powers does seem acutely aware that an idealistic attachment to the civilizing effects of the literary canon is past history. This could be of interest for the field of Law and Literature. To the extent that it advises lawyers to read a set of books claimed to sensitize the reader to the right kind of practical wisdom Law and Literature may be fighting a lost cause. Not because computer science is taking over (in the shape of the digital humanities) but because the claim that lawyers should all read Shakespeare's *Merchant of Venice* has long been challenged by those who seek altogether different stories, outside the canon, to give voice to altogether different conceptions of what should matter in law.[xviii] More interesting is what has been called Law *as* Literature (White 1990; Gaakeer 1998), which moves into the epistemological affinities between Law as a discipline that is involved with the ambiguity of texts and the need for judgement (and much more than that) and Literature as a discipline that is similarly and alternatively involved with the same matters (and much more than that). The difference is that law also deals with the interpretation of real life action and that its judgements cut into the flesh (life, liberty and property) of living persons. Law is violence (e.g. Cover 1995), in the end. As much as it aims to prevent, outsmart and replace violence.

For something to suffer from violence embodiment and situatedness seem preconditional. Even if the monopoly on violence of today's liberal democratic state is reasonably abstract in comparison to the era of torture and corporal punishment, the threat to one's liberty and property in the name of the law is for real. And such a threat would be lost on a system that cannot feel pain, humiliation, deprivation, discrimination, invasion or restriction of movement. Powers' Helen is an impossible event. She comes into being as a person with affections, a growing sense of beauty and ends up with regrets. When she realises what is missing she laments her disembodiment. She is worse of than a brain in a vat, because she is not even a brain. To all our knowledge, a thing-person like Helen is not going to happen. The novel would not be convincing if Powers were trying to

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

present us with science *fact*ion; a program cannot understand that it cannot understand. Not in our terms, that is. But Powers is making another point. He is showing what machine learning and neural networks afford, how clever and sophisticated they have become and how easily we may be fooled. His novel is a prophecy about what is in line for us, if we continue on this road. Programs that feed on the data we stack together will allow us to see with new eyes what we took for granted. They will surprise us by deciphering implicit wit, projection, sorrow, and many of the hidden associations in the use of language and literature. Or they will show us our arrogance, disinterest, verbosity, and empty metaphors. The prophecy goes further, however, by demonstrating the embarrassing abstraction of pure syntax, the shallowness of a program that can only infer from what we have first compiled – without ever having a clue of the underlying meaning. It can, however, create new meaning, thanks to our efforts to interpret what it produces. While we invent these programs they reinvent us, as Ihde (2008) rightly observed.

THE LEGAL STATUS OF SMART CONTRAPTIONS: TOOLS, RIVALS OR COMPANIONS?

Embodiment, emotion and cognition

Helen's lack of feeling seems the crucial issue. Though one could say there is hope for Helen, because she seems to feel that she cannot feel, this paradox may be the weak spot of the novel. Only strong AI would permit us to think that an artificial brain in a vat can 'understand' what it lacks in terms of embodied experience. By now, cognitive science has discovered the central role of emotion in cognition, notably in decision making (Damasio 2000), and this has spilled over in AI research (Minsky 2006). The fact that Helen is not for 'real' when she becomes aware of what she cannot feel can also be seen as a strong point of the novel. It confronts the reader with a paradox, a tension, an impossibility, that invites further imagination, thought and discernment. The fact that Helen's self consciousness is fictional does not mean that artificial intelligent life forms cannot emerge. I would agree with a number of scholars that we cannot rule out that non-biological man-made contraptions will come alive (Bourgine en Varela 1992; Brooks 1991; Pfeifer en Bongard 2007),

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

though this does not necessarily imply consciousness, let alone self-consciousness.

Pioneering work on the nexus of cognitive psychology and computer science has been done by Picard (1995), under the heading of affective computing. Her aim has been to use computers to recognize and diagnose emotions and to further investigate the role they play in cognition. Some researchers even go so far as to develop what they call synthetic emotions (Velasquez 1998), to make machine decision more effective by programming machine readable versions of pain and pleasure into the software as sticks and carrots. Synthetic emotions, however, that are based on our own embodiment will not do for artificial life forms. Their emotions will have to emerge from their own experience as embodied entities instead of being imposed on them. One of the most daunting explorations of this position has been made by Pfeifer and Bongard (2007), who develop a sophisticated grounded theory of *How the body shapes the way we think,* claiming that by attempting to build systems that can develop into what they call 'complete agents' we may discover some of the misconceptions we have about our own mind. There are drawbacks here. As Picard noted in 1995, there may be a risk in building machines with emotions, for we cannot take for granted that they will care for us in a way that contributes to human flourishing. To the extent that emotion is connected with survival, as Damasio (2000) and many other psychologists claim, these machines may become our rivals, adversaries or even enemies at some point in the future. They will probably not compete for a master in English literature, but be built as agents to improve profits, police investigation or scientific research.

Some thoughts on the legal implications of smart agents

In other work we have traced some of the implications of the rise of artificial agents for the notion of legal personhood in private and in criminal law (Koops, Hildebrandt, en Jacquet-Chiffelle 2010; Hildebrandt 2011). The first legal scholar to make an original and comprehensive analysis of the issue was Solum (1992), who decided to evade the metaphysical question of 'what is intelligence' and to replace it with the pragmatic issue of whether an AI could take on

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

the legal role of a trustee. One could see this as a lawyer's version of the Turing Test. Solum's main practical point was that AIs were not (yet?) capable of judgements that require a measure of discretion, even though they might be able to take over a number of decisions that merely require the straightforward application of straightforward rules to straightforward cases. This is a well-known argument in the literature on legal knowledge systems which are used as tools for the automated implementation of legal rules. Such systems are presently employed to 'process' decisions on social welfare, traffic fines, taxes, and other types of administrative decisions that involve massive amounts of routine decisions. Most authors agree that the real problem here is that the question of whether a case is straightforward (easy) or complex (hard) is itself a question that cannot be answered by the system, because it requires the kind of discernment, discretion and judgement they lacked in the first place. Notably Leenes (1998), Van der Linden-Smith (2000), Citron (2007) have discussed these issues in depth. Solum extended his analysis with a different question, by asking whether AIs should be granted constitutional protection. Though the answer to the first question mostly concerns breach of contract or tort liability, a positive answer to the second question would in fact attribute life, liberty and property to AIs. Solum was of the opinion that in the end the question of whether AIs should be granted legal personhood is an empirical question, depending on the legal role they should play and on the extent to which they can actually fulfil this role. Since 1992 interesting work has been done, taking into account the ephemeral, polymorphous and mobile character of artificial agents (e.g. Karnow 1997) and or advancing the perspective of legal theory (e.g. Chopra & White 2011).

My aim here is not to develop a set of conditions to be fulfilled by an artificial agent for us to grant it legal personhood. Instead I want to cherry-pick two types of questions that we should confront while developing smart computing systems like Helen. They both relate to the notion of 'agency'. Since this position paper is already so much longer than permitted I will restrict myself to a brief indication of their scope and hope for an interesting discussion in Frankfurt this August.

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

The notion of legal personhood for artificial intelligence is often connected with the notion of 'agency'. The reason is twofold.

I On the one hand moral philosophy uses the term human agency to refer to the assumption that human beings act on the basis of intentions and can give reasons for their actions. Entities which lack such intentional states cannot be held responsible for their behaviours. Think of vulcanos, nuclear plants or webbots. Either no liability can be attributed (in the case of an Act of God that could not be foreseen), or human beings or organisations are held liable for designing, producing or using the entity. The legal status of an entity without this type of agency is that of a tool.

II On the other hand, a more mundane meaning of agent refers to a legal person who acts in the name of and/or on behalf of a patron. This legal figure enables the patron to act through his agent, meaning that the agent can – if certain conditions are fulfilled - legally bind the patron to a contract concluded with a third party. The agent is some kind of intermediary. Within computer science artificial agents are often used to fulfil well defined tasks for its user (e.g. a webbot that searches the web for certain information, buys books, airline tickets or whatever). It has been noted that for a computer agent to qualify as a legal agent it would need legal personhood.

Both meanings of 'agency' raise questions as to the desirability of legal personhood for bots.

I As to the question of legal personhood for intermediaries: does and if so, should our legal system allow us to provide a limited kind of legal personhood to artificial agents? what advantages would this bring for their patrons and for those they engage with as parties to a contract or as a party having committed a specific tort? how does legal personhood for programs or machines relate to legal personhood for corporations or animals?

II As to the question of human agency in the sense of having the capacity for reasons and intentions: would it be reasonable to regard embodied agents that develop some kind of agency in the philosophical sense of the word as mere agents (intermediaries) that

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

should serve our purposes only? could 'they' claim entitlement for human rights protection and whatever could this mean for a nonhuman? would the integration of synthetic emotions entitle artificial agents to human rights protection? what if 'they' claim political rights and citizenship?

BIBLIOGRAPHY

Black, Edwin. 2002. *IBM and the Holocaust. The Strategic Alliance between Nazi Germany and America's Most Powerful Corporation*. New York: Crown. http://www.ibmandtheholocaust.com/.

Bourgine, Paul, en Francisco J. Varela. 1992. Towars a Practice of Autonomous Systems. In *Towards a Practice of Autonomous Systems. Proceedings of the First European Conference on Artificial Life*, bewerkt door. Francisco J. Varela en Paul Bourgine, xi-xviii. Cambridge, MA: MIT Press.

Brooks, Rodney. 1991. Intelligence without Reason. In Proceedings of the Twelfth International Joint Conference on Artificial Intelligence, 569-595.

Chopra, Samir, en Laurence F. White. 2011. A Legal Theory for Autonomous Artificial Agents. University of Michigan Press.

Christian, Brian. 2011. The Most Human Human. What Talking with Computers Teaches Us About What It Means to Be Alive. New York: Doubleday.

Citron, Danielle K. 2007. Technological Due Process. *Washington University Law Review* 85, no. available at: http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1012360: 1249-1313.

Cohen, Robert. 1995. Pygmalion in the Computer Lab, July 23.

Cole, David. The Chinese Room Argument. In: Edward N. Zalta. *The Stanford Encyclopedia of Philosophy*. Winter 2009 Edition: http://plato.stanford.edu/archives/win2009/entries/chinese-room/.

Cover, Robert, Martha Minow, Michael Ryan, en Austin Sarat. 1995. *Narrative, Violence, and the Law. The Essays of Robert Cover*. Michigan: University of Michigan Press.

Custers, Bart. 2004. *The Power of Knowledge. Ethical, Legal, and Technological Aspects of Data Mining and Group Profiling in Epidemiology*. Nijmegen: Wolf Legal Publishers.

Damasio, Antonio R. 2000. *The feeling of what happens : body and emotion in the making of consciousness*. New York: Harcourt Inc.

Dreyfus, Hubert L. 1979. *What computers can't do : the limits of artificial intelligence*. Harper colophon books CN 613. New York: Harper & Row.

———. 1992. *What computers still can't do : a critique of artificial reason*. Cambridge, Mass. ; London: MIT Press.

Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth, en Ramasamy Uthurusamy. 1996. *Advances in Knowledge Discovery and Data Mining*. Meno Park, California - Cambridge, Mass. - London England: AAAI Press / MIT Press.

Floridi, Lucia, en Mariarosaria Taddeo. 2009. Turing's Imitation Game: Still an Impossble Challenge for All Machines and Some Judges - An Evaluation of the 2008 Loebner Contest. *Mind and Machines* 19, no. 1: 145-150.

Gaakeer, Jeanne. 1998. *Hope springs eternal: An introduction to the work of James Boyd White*. Amsterdam: Amsterdam University Press.

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

Hildebrandt, M. 2012. Eccentric positionality as a precondition of the criminal liability of artificial life forms. In *Artificial by Nature. Plessner's Philosophical Anthropology. Perspectives and Prospects* Maarten Coolen, Huib Ernste, en Jos De Mul (eds.), Amsterdam: Amsterdam University Press 2012

———. 2011a. Criminal liability and "smart" environments. In *Philosophical Foundations of Criminal Law*, bewerkt door. Antony Duff en Stuart Green, 507-532. Oxford: Oxford University Press.

Hildebrandt, M, en Antoinette Rouvroy. 2011b. *Law, human agency and autonomic computing. The philosophy of law meets the philosophy of technology*. Abingdon: Routledge.

Hildebrandt, M., en Serge Gutwirth. 2008. *Profiling the European Citizen. Cross-disciplinary Perspectives*. Dordrecht: Springer.

Hildebrandt, Mireille. 2011. Autonomic and autonomous "thinking": preconditions for criminal accountability. In *Law, Human Agency and Autonomic Computing*. Abingdon: Routledge.

IBM White Paper. 2011. Watson - A System Designed for Answers. The future of workload optimization.

Ihde, Don. 1991. *Instrumental realism : the interface between philosophy of science and philosophy of technology*. The Indiana series in the philosophy of technology. Bloomington: Indiana University Press.

———. 2008. *Ironic Technics*. Automatic Press.

Karnow, C.E.A. 1997. *Future Codes: Essays in Advanced Computer Technology and the Law*. Boston London: Artech House.

Koops, B.J., M Hildebrandt, en David-Olivier Jacquet-Chiffelle. 2010. Bridging the Accountability Gap: Rights for New Entities in the Information Society? *Minnesota Journal of Law Science & Technology* 11, no. 2: 497-561.

Kranzberg, Melvin. 1986. Technology and History: "Kranzberg"s Laws'. *Technology and Culture* 27: 544-560.

Kurzweil, Ray. 2005. *The singularity is near : when humans transcend biology*. New York: Viking.

Leenes, R. 1998. *Hercules of Karneades. Hard cases in recht en rechtsinformatica*. Enschede: Twente University Press.

Van der Linden-Smith, Tina. 2000. Een duidelijk geval: geautomatiseerde afhandeling.

Markoff, John. 2011. Computer Wins on "Jeopardy": Trivial, It's Not, februari 16.

Minsky, Marvin. 1988. *The Society of Mind*. Simon & Schuster.

Minsky, Marvin. 2006. The Emotion Machine. Simon & Schuster.

Mitchell, Tom M. 2006. The Discipline of Machine Learning. Carnegie Mellon University, School of Computer Science, available at http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf.

Moore, Gordon E. 1965. Cramming more components onto integrated circuits. *Electronics Magazine*.

Pfeifer, Rolf, en Josh Bongard. 2007. *How the Body Shapes the Way We Think. A New View of Intelligence*. Cambridge, MA - London, England: MIT Press.

Picard, Rosalind. 1995. *Affective Computing*. Cambridge, MA: MIT.

Powers, Richard. 1995. *Galatea 2.2*. New York: Picador.

Powers, Richard. 2011. What Is Artificial Intelilgence? *The New York Times*, februari 5.

Searle, John. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3, no. 3: 517-557.

Shannon, Claude, E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*.

Simon, Herbert A. 1996. The Sciences of the Artificial. Cambridge MA: MIT Press.

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

Solum, Lawrence B. 1992. Legal Personhood for Artificial Intelligences. *North Carolina Law Review* 70, no. 2: 1231-1287.

Steels, Luc. 1995. When are robots intelligent autonomous agents? *Robotics and Autonomous Systems* 15: 3-9.

Turing, A.M. 1950. Computing Machinery and Intelligence. *Mind* 49: 433-460.

Velasquez, J.D. 1998. Modeling Emotion-Based Decision Making. In *Emotional and Intelligent: The Tangled Knot of Cognition*, 164-169.

Wiener, Norbert. 1948. *Cybernetics: Or Control and Communication in the Animal and the Machine*. Cambridge MA: MIT Press.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman \& Co.

White, James Boyd. 1990. *Justice as translation: an essay in cultural and legal criticism*. Chicago: University of Chicago Press. http://mirlyn.lib.umich.edu/Record/002168401.

---

[i] This paper builds on (Mireille Hildebrandt 2011a), (M Hildebrandt 2011b), (M Hildebrandt 2012.

[ii] Naming it Helen does remind one of Trojan horses, a nice overlap between world literature and computer science.

[iii] See (Dreyfus 1979) and again (Dreyfus 1992) for a sustained critique from the perspective of phenomenology. His work had a major influence in the field.

[iv] A remarkable attempt to link the fundamental uncertainties uncovered by the natural sciences with the humanities was made by Prigogyne and Stengers in their well known discussion of chaos theory. The original French title of their book was *La nouvelle alliance. Metamorphose de la science (1979).*

[v] Turing's 1950 article is a very sophisticated and unorthodox exploration of what he calls 'the imitation game'. Many of the objections that have been made since then are already foreseen and countered by Turing in this article. The point is not whether one agrees, but to detect to what extent his predictions have come true. See (Floridi en Taddeo 2009) for an evaluation of the 2008 Loebner Contest, a yearly event that imitates the Turing Test and nominates 'the most human machine' as wel as 'the most human human'. See Christian (2011) who played as human in the 2009 Loebner Contest and came out as 'most human hutman'.

[vi] See Christian (2011), chapter 7 'Barging in' on the silliness as well as the rigidity of much chatbots' conversation.

[vii] See Christian (2011), chapter 5 'Getting Out of Book' on the reliance on registered games.

[viii] Futurist Kurzweil (2005) has coined the term singularity for the moment in time when all problems that are intractable now will be resolved. This will be the moment that 'humans transcend biology'. Only those who believe that all problems that matter are

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.

computable will be relieved to hear this. My point is that even if all problems are computable they are usually computable in different ways, with different outcomes. Back to square one?

[ix] This kind of robust knowledge, however, requires transparency as to the translations, requiring access to the whole process of knowledge construction. This is not possible as long as this kind of knowledge production is protected by trade secret and/or intellectual property rights.

[x] (Black 2002). As we know, you can use a knife to slice beef or to kill your fellow; though a technology in itself is neither good nor bad, it is never neutral (Kranzberg 1986).

[xi] See (IBM White Paper 2011): To achieve the most right answers (in the case of Jeopardy: the most right questions) at a competitive speed, IBM deploys: (1) massive parallelism to consider multiple interpretations and hypothesis; (2) many different experts to integrate, apply and contextually evaluate loosely coupled probabilistic questions with content analysis; (3) confidence estimation on the basis of a range of combined scores; and finally (4) integration deep and shallow knowledge, leveraging many loosely formed ontologies.

[xii] Data science is 'the new kid on the block'. It provides a set of tools to infer knowledge from Big Data and is used in all the sciences now, from the natural sciences, to the life sciences, to medicin and healthcare, the humanities and the social sciences. Plus marketing and customer relationship management, forensic science and police intelligence. See notably Mitchell (2006); Fayyad e.a. (1996); Custers (2004); M. Hildebrandt en Gutwirth (2008).

[xiii] In fact, in the case of the game of Jeopardy, Watson has to find precise questions to specific answers.

[xiv] Moore (1965), Intel co-founder, predicted that the computing power of chips would increase exponentially (doubling every two years). The prediction became a goal for the industry which has so far been met.

[xv] The addition of 2.2 to Galatea seems to refer to version 2.2 of the program that constitutes Helen.

[xvi] For a more extensive discussion see Cole (2009).

[xvii] Shaw's (1902) *Pygmalion* was the inspiration of the romantic musical *My Fair Lady* ( ). Note that Galatea translates as 'she who is white as milk', which seems a 'fair' translation of Shaw's Eliza Doolittle and remember that Weizenbaum's therapeutic machine was called Eliza.

[xviii] This is – evidently – not to discredit Shakespeare or the *Merchant of Venice*. It is to say that we cannot take for granted what is relevant and should not too easily think in terms of a canon.

Please cite from: Mireille Hildebrandt, From Galatea 2.2 to Watson – and Back?, in: M. Hildebrandt and J. Gaakeer (eds.), *Human Law and Computer Law: Comparative Perspectives*, Springer 2013, 23-45, http://dx.doi.org/10.1007/978-94-007-6314-2_2.