Philosophy & Technology (2021) 34:1195–1211
https://doi.org/10.1007/s13347-021-00453-8

RESEARCH ARTICLE

Computing Machinery, Surprise and Originality

Sylvie Delacroix[1]

Received: 9 March 2021 / Accepted: 4 May 2021 / Published online: 21 May 2021

Abstract
Lady Lovelace's notes on Babbage's Analytical Engine (1843) never refer to the concept of surprise. Having some pretension to 'originate' something—unlike the Analytical Engine—is neither necessary nor sufficient to being able to surprise someone. Turing nevertheless translates Lovelace's 'this machine is incapable of originating something' in terms of a hypothetical 'computers cannot take us by surprise' objection to the idea that machines may be deemed capable of thinking. To understand the contemporary significance of what is missed in Turing's 'surprise' translation of Lovelace's insight, one needs to distinguish between trivial surprises (which stem from our limited ability to store data and process it) and those events, propositions or encounters that lead us to question our understanding of ourselves or what surrounds us. Only some of these non-trivial surprises are the product of originality endeavours. Not only is it uncommon for surprises to track such endeavours, the type of autonomy that would be required on the part of 'digital computers' for originality and surprise to intersect in that way goes far beyond the operational autonomy that can be achieved by 'learning machines'. This paper argues that a salient translation of Lovelace's originality insight—for contemporary and future 'learning machines'—is an upside-down version of Turing's surprise question: can computers be surprised by us in a non-trivial, 'co-produced' way?

Keywords Turing · Lady Lovelace · Surprise · Originality · Interpretive capabilities · Hermeneutics · AlphaGo

1 Introduction

We have devised machines that help us keep track of time, cultivate the earth, mend our bodies, survey the skies…the list goes on. Some aim to overcome specific physical limitations; other machines are designed to entertain. Many do both. Most have

✉ Sylvie Delacroix
sdelacroix@turing.ac.uk

[1] University of Birmingham & Alan Turing Institute, London, UK

had a profound impact on our understanding of the world, and the role we can play within it. None more so than one of our more recent inventions: computers.[1]

In that context, to ask whether computers are able to 'take us by surprise' may sound like a redundant question. When it is famously raised by Alan Turing (Turing, 1950), nobody is in a position to predict the depth and extent of the socio-cultural upheavals brought about by their near-universal use today. Yet, this historical upheaval is not what Turing has in mind when he floats the 'computers cannot take us by surprise' proposition, only to dismiss it as unsubstantiated. Turing points out that computers do take him by surprise all the time, given their ability to fill the gaps in his incomplete (or deficient) calculations.

This 'computers cannot take us by surprise' idea is raised by Turing in the context of his enquiry into whether machines can think. To dis-ambiguate this inquiry, Turing proposes to replace the 'can machines think?' with a different question: are there imaginable, 'digital'[2] computers that would do well in an 'imitation game'? The purpose of that game is to fool a human observer, whose aim is to tell the computer from the human in a question and answer game. Among the arguments that may be raised to dispute the idea that computers could ever do well at that game, one may point at various 'disabilities' of computers. Having surveyed the latter, Turing considers whether Lady Lovelace raises a different sort of argument when, describing Babbage's 'Analytical Engine' (in 1843), she noted that the latter 'has no pretension to *originate* anything. It can do *whatever we know how to order it* to perform' (Turing, 1950, p. 12). Turing first translating the above insight—about the (in)ability to *originate* something—in terms of an (in)ability to surprise him is peculiar. Interestingly, Turing switches to a rather different take on Lovelace's originality insight when, towards the end of his paper, he suggests that 'learning machines' could be another way of responding to her 'computers lack the capacity to originate' objection.

This learning-focused response had the potential to bring to light the relationship between different learning processes and the kinds of autonomy enabled by such processes. The learning required to be capable of creative autonomy is different from that which underpins 'mere' operational autonomy. The relationship between these kinds of autonomy and the sorts of surprises they may generate is unpacked in §2, which distinguishes between first-, second- and third-person accounts of originality endeavours. Whereas third-person accounts tend to define originality in terms of purely epistemic criteria (such as whether a given output could have been anticipated), first-person accounts consider the capabilities and needs underlying originality endeavours. To understand the latter, one needs to differentiate between merely original outputs (a spontaneous toddler dance) and

[1] In the rest of this paper, the word 'computer' is meant to be understood loosely—as any device that can store and process information: thus, it is used to refer to machines such as Babbages' 'Analytical Engine' all the way to speculative, autonomous artificial agents. The only meaning of 'computer' that is not included under this very loose definition is that which is used to characterise humans performing calculations.

[2] 'The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer' (Turing, 1950).

one's purported ability to 'originate': one has not 'originated' anything if one's 'original' output does not have any future. The latter, diachronic aspect entails second-person accounts of originality. Such accounts are key to delineating the relationship between originality and surprise, which only occasionally overlap.

In order to articulate this relationship, §3 starts by unveiling the ambiguity inherent in the term 'surprise' and draws two distinctions. Turing's pointing out that his hurried calculations are frequently proven wrong, in a way that surprises him, highlights the importance of first distinguishing between trivial and non-trivial surprises: the term 'surprise' can be used quite prosaically, to refer to any event or 'input' that we fail to anticipate. In that mundane sense, we cannot but be surprised all the time, given our limited ability to store data and process it. Some surprises, on the other hand, are less trivial: they come closer to what Fisher describes as an experience of 'wonder', which may lead us to question our understanding of ourselves or what surrounds us (Fisher, 1998).

These non-trivial surprises can stem from a variety of things, from rainbows to artworks, via human encounters. Only some of these are the product of originality endeavours. This second distinction—between surprises that can be said to be 'co-produced', in that they are made possible in part by some originality endeavour, and those that are not—helps delineate what is missed in Turing's 'surprise' translation of Lovelace's originality insight. Not only is it uncommon for surprises to track originality endeavours, the type of autonomy that would be required on the part of 'digital computers' for originality and surprise to intersect in that way goes far beyond the operational autonomy that can be achieved by 'learning machines' (whether Turing's or today's).

Building upon the double distinction introduced in §3, §4 considers contemporary endeavours to optimise a system's learning capacity on the basis of so-called 'surprise modulated belief algorithms'. As a tool to translate the need for a minimal degree of model uncertainty if a system is to keep adequately learning from its environment (hence retaining the ability to be surprised), a narrow, quantifiable concept of surprise is fairly harmless. Yet, when such 'surprise modulated belief algorithms' are presented as potential ways of modelling the way *humans* learn within dynamic environments, things start getting both problematic and interesting, given the magnitude of what is missed (or dismissed). It is precisely because human learning processes also aim at creative—rather than merely operational—autonomy that an account of those processes in terms of error-minimisation imperatives is unhelpful.

§5 concludes by re-considering the pertinence of Turing's discussion of Lovelace's insight in the context of his 'can machines think' discussion. Since our current 'digital machinery' arguably lacks the interpretive capabilities that underpin endeavours to originate something, Lady Lovelace is proven right. This does not make Turing's question any less pertinent, especially if one is prepared to consider it upside down. The most helpful translation of Lovelace's originality insight turns out to be a reversed version of Turing's surprise question: can computers be surprised *by us* in a non-trivial, 'co-produced' way?

2 From Third-Person to First- and Second-Person Accounts of Originality

Lovelace's Notes on Babbage's Analytical Engine not only included a fuller description of the way the Engine worked (compared to Menabrea's French-language account). They also analysed the Engine's potential and limitations, among which the passage quoted in (Turing, 1950):

> The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform… (note G). (Lovelace, 1843)

Turing's translating Lovelace's insight—that Engine's (in)ability to *originate* something—in terms of an (in)ability to surprise him (or humans generally) is intriguing. Originality and surprise are distinct concepts which overlap only occasionally. I may deploy novel methods to generate outputs that challenge accepted norms but fail to surprise anyone (as is the case of many unrecognised artists), and conversely: my outputs may have been produced by following to the letter a set of instructions, and yet they may surprise many (the set of instructions may have been designed to maximise the chances of triggering 'surprise' in my audience).

Peculiar as it is, Turing's 'originality as capacity to surprise' translation is significant: it reflects methodological assumptions that still shape much of today's cognitive sciences, including debates on artificial agency. This section highlights the contrast between these assumptions (unpacked in §2.1.) and the types of questions that arise from first- and second-person accounts of originality (§2.2. and §2.3.).

2.1 Third-Person Accounts of Originality

The year immediately following the publication of his 1950 'Computing Machinery and Intelligence', Turing gives a broadcast which focuses almost exclusively on Lovelace's originality insight (Turing, 2004). This BBC broadcast helpfully spells out what I shall call the 'downstream epistemic' understanding of originality:

> If we give the machine a programme which results in its doing something interesting which we had not anticipated, I should be inclined to say that the machine had originated something, rather than to claim that its behaviour was implicit in the programme, and therefore that the originality lies entirely with us. (Turing, 2004, p. 485)

Such an account is 'downstream' in that the originality assessment is wholly dependent upon the recipient's characterisation. The criteria at play are also overwhelmingly epistemic: rather than dwell on when something might be deemed 'interesting', Turing focuses on the anticipation criterion, or what is sometimes called the 'epistemic limitation' in the contemporary literature: 'Satisfaction of the epistemic–limitation condition by a machine amounts to performance by the

machine in ways unforeseen by its creator (or someone with access to the resources its creator did)' (Abramson, 2008, p. 160).[3] On this account, originality would be conditional upon the impossibility of anticipating the agent's behaviour, given what we know about the machine's (or human's) 'initial state, and input'. Since such a definition departs from our everyday understanding of originality, one may seek to query the reasons behind it (if an attentive mother finds it impossible to anticipate the everyday behaviour of her toddler, the latter is not necessarily deemed 'original').

The rationale behind this 'epistemic' delineation of originality may be reconstructed along the following lines: without presupposing what needs to be demonstrated—the possibility of machine agency—some sort of 'thinking' autonomy could only be inferred from epistemic limitation conditions. The problem with this line of reasoning is that it allows methodological constraints to overly narrow down the definition of what needs to be demonstrated: autonomous agency. The latter comes in different kinds and degrees. While a pupil who has mastered long divisions may be said to have 'operational' autonomy in that respect (just like a computer that can play chess), creative autonomy presupposes the ability to question and change given rules or conventions (Boden, 1998). The latter is also different from what may be referred to as 'fantastic' autonomy: it is much more difficult to create something new from within a background of accepted norms and conventions (i.e. creative autonomy) than to do so in the absence of any prior constraint or interpretive convention. Such 'ex-nihilo' creativity may be called 'fantastic' in that it is not the way we humans generate novelty. Yet, the third-person, epistemic perspectives on originality sometimes seem to presuppose this 'radically unprecedented' sort of autonomy.

These third-person perspectives are also unable to capture the diachronic dimension inherent in the term 'originate'. While some inadvertent artistic output may be deemed original, for the artist to be deemed to have originated something, that output needs to have some future: unless it prompts the emergence of some intellectual, artistic or socio-cultural practice, that output's originality only captures its fleeting novelty in relation to existing practices. Whether an output's saliency perdures beyond that fleeting assessment—in which case the output's creator has 'originated' something—depends on both the qualities of that output (some posit the latter as conditions of 'objective'[4] or 'inter-subjective'[5] creativity'[6]) and its audience's receptivity.

[3] See also Bringsjord et al. (2001) for a similarly 'downstream epistemic' focus.

[4] Jarvie (1981, p. 117) distinguishes between subjective creativity, which is 'a property of persons or their minds', and objective creativity, which is 'a property […] of created works'.

[5] Inter-subjective creativity is an attribute of artifacts, whereas subjective creativity is an attribute of acts (D'Agostino 1986, p. 175).

[6] Along this line, Gaut (2003, p. 270) argues that 'Creativity in the narrower non-modal sense is the kind of making that involves flair in producing something which is original (saliently new) and which has considerable value.'.

2.2 First-Person Accounts of Originality

Rather than focusing solely on what counts as 'original', a first-person perspective will seek to understand what drives originality as an effort. What needs or desires does a quest for originality seek to fulfil? What, if anything, might compromise a capacity to originate something? Might Lovelace's '*cannot originate anything*' verdict ever apply to human agents? These questions fit within a wider, long-standing philosophical endeavour to understand human agency.

Our interactions with the world around us are structured by beliefs, norms or desires that are shaped by our socio-historical circumstances. Today, a large proportion of philosophers will readily acknowledge this contingency,[7] even if they will differ in their articulation of its epistemic and normative implications. Putting aside nihilistic, 'anything goes' accounts (whether in epistemology or ethics), one step forward is to de-dramatise this acknowledged contingency:

> Because we are not unencumbered intelligences selecting in principle among all possible outlooks, we can accept that this outlook is ours just because of the history that has made it ours; or, more precisely, has both made us, and made the outlook as something that is ours. (Williams, 2000, p. 491)

To be genuinely relaxed about the latter conclusion is not without its challenges: from an epistemological perspective, it requires abandoning any striving to know the world as it is in itself (independently of our representations of it). From an ethical perspective, one must be able to account for the possibility of change: that some contingent history has made us so does not mean we cannot set in motion a chain of events that will change its trajectory. It is less easy to show how we do that, if all we have to trigger the movement of critical scrutiny are our culturally conditioned habits of evaluation, rather than some Archimedean point safely removed from the contingent mess of human affairs.

Whether in the domain of aesthetics, epistemology or ethics, it is only once one takes on board the 'encumbered' nature of our intelligence that the full significance of originality as an imperative—rather than nicety—comes into light. If that is so, some may wonder why originality should be of any relevance as an indicator of *artificial* intelligence? If what drives originality as an effort (and perhaps a sign of intelligence) is the need to be able to challenge and enrich the web of socio-cultural expectations that shapes us from birth, then there must be room for sophisticated learning machines that have no such need for originality. Unlike humans, surely such machines could be in a position to control what they let themselves be 'shaped' by?

In its pragmatism, the above rejoinder however confuses needs and imperatives, and in doing so misses the heuristic value of Lovelace's insight, which has only grown since Turing first sought to capture it. For better or worse, many efforts to

[7] For a paper brilliantly outlining why there is no plausible way of vindicating the critical genealogist's inference from alethic indifference ('our beliefs are said to be produced by a causal mechanism that we have no independent reason to believe will tend to produce true beliefs') to lack of knowledge, see Srinivasan (2019).

build learning agents are premised upon the validity (and helpfulness) of a (young) human–machine analogy: 'We believe that to truly build intelligent agents, they must do as children do' (Kosoy et al., 2020, p. 4). So far this analogy has been interpreted narrowly. Efforts to understand children's learning processes often proceed from the assumption that these learning processes are mostly aimed at building a model of the world that minimises prediction errors. A good fit for many problems involving perception and motor control, the latter goal certainly makes for smoother, more efficient interactions with one's environment. Yet, it arguably 'leaves out much that really matters for adaptive success'.[8]

To understand the value of other goals like play—or originality—an evolutionary narrative (of the type associated with the prediction error minimisation framework[9]) will only go so far if it is not associated with a study of socio-cultural imperatives. At an individual level, originality as a goal is likely entangled with identity formation (as a process that tends to take a more dramatic turn in teenage years[10]). At a collective level, endeavours to originate something will be tied to the mechanisms that underlie socio-moral change. This collective aspect calls for a second-person account of originality.

2.3 Second-Person Accounts of Originality

Few twentieth century philosophers can claim to have been more preoccupied by the factors that may compromise our capacity for 'new beginnings'[11] than Hannah Arendt. Animated by a concern to see us retain our capacity for political action, which 'marks the start of something, begins something new', Arendt feared the consequences of its atrophy under the weight of everyday 'thoughtlessness' (Arendt, 2007, p. 113).[12] At stake is not just the possibility of ethical or political change ('viewed objectively and from outside, the odds in favour of tomorrow unfolding just like today are always overwhelming') (Arendt, 2007, p. 112). The loss of our capacity for originality would also undermine what is distinctive, and perhaps dignified,

[8] Clark (2013, pp. 12-13) raises this in an attempt to address the 'very general worry that is sometimes raised in connection with the large-scale claim that cortical processing fundamentally aims to minimize prediction error'.

[9] 'Change, motion, exploration, and search are themselves valuable for creatures living in worlds where resources are unevenly spread and new threats and opportunities continuously arise. This means that change, motion, exploration, and search themselves become predicted—and poised to enslave action and perception accordingly' (Clark, 2013, p. 13).

[10] From an evolutionary perspective, one can imagine narratives that tie originality as a goal to the value inherent in variance.

[11] In *The Human Condition*, Arendt (1998) draws an important distinction, based on St Augustine, between 'the beginning which is man (*initium*)' and 'the beginning of the world (*principium*)'. As a way of illustrating the importance of this distinction, Arendt for instance argues that if the French revolutionaries had not understood their task as an absolute, godlike beginning (*principium*, beginning of the world), which is by definition beyond their capacities (and required a new calendar), they would probably have been able to avoid many of the perplexities (and ensuing violence) they were confronted with.

[12] Arendt (1971, p. 445) refers to the peril inherent in those times 'when everybody is swept away unthinkingly by what everybody else does and believes in'.

about the way we are always in the process of constructing human selfhood. Never settled in advance, 'we are not anything definite until we reveal our "who" [to others] in speech or deed' (Hinchman & Hinchman, 1984, p. 202).

Because the beliefs and norms that preside over the receptivity of those 'others' are themselves shaped by contingent circumstances, each endeavour to originate something presupposes the judicious combination of both 'reproductive' and 'productive'[13] imagination (Ricoeur, 1975):

> Any transformative fiction – any utopia, any scientific model, any poem – must have elements of reproductive imagination, must draw from existing reality sufficiently so that its productive distance is not too great. (Taylor, 2006, p. 98)

It is this emphasis on originality's dependence upon the interpretive articulation of what Turner calls the 'unoriginal structures that inform originality'[14] that matters for our purposes. The difference between 'mere' output novelty and efforts to originate something cannot be grasped unless one acknowledges this second-person aspect of originality. As a hermeneutic effort, originality is not dependent upon success: what matters is that there be an endeavour to judiciously combine the productive and reproductive imagination(s) referred to earlier. Without this second-person dimension, originality could be solely a matter of productive imagination. While the latter could be argued to be within the reach of today's machines,[15] the interpretive capabilities required for reproductive imagination are still peculiarly human.[16] Arendt is far from the only philosopher to emphasise the importance of these interpretive capabilities,[17] yet she articulates them in a way that brings home what is strikingly absent in both Turing's and contemporary emphases on unforeseeability as a (downstream) criterion for originality.

In contrast to such unforeseeability criterion, Arendt's account highlights precisely the opposite:

> The very originality of the artist (or the very novelty of the actor) depends on his making himself understood by those who are not artist (or actors). (Arendt, 1989)

Arendt's emphasis on the second-person perspective is crucial: to be capable of originating something requires interpreting[18] a shared web of socio-cultural

[13] Referred to in Taylor (2006).

[14] 'Originality, far from being autonomous, is contingent at every point upon the unoriginal structures that inform it' (Turner, 1991, p. 51).

[15] See §5.

[16] These interpretive capabilities are sometimes referred to as 'a degree of understanding'. See for instance: 'if, as I will argue shortly, creative activity requires some degree of understanding, then, if computers cannot exhibit understanding (Searle), they cannot be creative' (Gaut, 2010).

[17] One may speculate about what might have happened had Turing and Arendt met and discussed (an unlikely proposition, despite being both already world-renowned in their respective fields). Chronologically, Arendt presented an early version of her Kant Lectures in Chicago in 1964, and was likely working on the material leading to these lectures a mere decade after Turing's BBC podcast.

[18] This interpretation need not be deliberate: 'not all radically creative acts involve deliberate attempts to transform conceptual spaces' (Novitz, 1999, p. 72).

expectations in such a way as to remain intelligible[19] while at the same time challenging those expectations. This challenge may result in an experience of novelty and/or surprise. The next section articulates the relationship between originality endeavours and surprise, which only occasionally overlap (and which third-person perspectives unhelpfully blur).

3 The Relationship Between Kinds of Surprises and Originality

Consider two contemporary uses of machine learning techniques, and the kind of surprises they might generate:

1. A system trained to explore possible configuration spaces and predict the atomic structure of crystals (according to the equations of quantum mechanics) will come up with lots of atomic structures, some of which might be stable enough, energy-wise, to constitute a 'plausible' crystal (Deringer et al., 2018). Aside from filling the gaps in our incomplete knowledge, the generation of such atomic structures could also surprise scientists in a non-trivial way if it leads them to reconsider some of their hypotheses about the properties of some crystals.
2. When AlphaGo came up with a 'new' move (Silver et al., 2016), one that had never been considered before, did it 'originate' anything? The move itself was merely one of the x-many moves compliant with the rules of Go. Its significance stemmed from its challenging strategic assumptions widely shared among Go experts. The extent of AlphaGo's operational autonomy (which stemmed from a sophisticated learning process) combined with the vast search space (something like 10^170) increased its ability to challenge the expectations of even the most learned Go experts. None of them had anticipated the value of 'move 37'. This anticipation failure forced Go experts to reconsider their understanding of the game. In that sense, it was a 'generative' move, not a move that should count as 'original': it only required operational autonomy, and none of the hermeneutic effort described earlier. Were other members of the public surprised by this new move itself? No. If they were surprised, it was by that system's ability to surprise human experts: this forced them to reconsider their understanding of what 'digital machinery' could do.

Both of the above machine learning systems merely 'make available what we are already acquainted with', to use Lovelace's own words (Lovelace, 1843). We—or at least experts—are well acquainted with both the equations of quantum mechanics and the rules of Go. Those systems thus 'make available' atomic structures or moves, within the established realm of possible structures or moves.

[19] See also 'Creativity involves coming up with something novel, something different. And this new idea, in order to be interesting, must be intelligible' (Boden, 2010, p. 164).

To refer back to the two distinctions mentioned in the introduction: the surprises that stem from those structures or moves are not 'co-produced' in that they are not partly made possible by an originality endeavour at the upstream. To the extent that such structures or moves lead experts to re-consider the assumptions underlying their understanding of crystals (or Go), those surprises may also be said to be non-trivial.

Neither humans nor algorithms are capable of 100% accurate predictions in real-life applications. Prediction (or calculation) failures are expected. Yet some events or propositions (whether they were anticipated or not) do give rise to surprises in a way that goes beyond the 'banal',[20] 'unanticipated' sense. They do so when they lead us to reconsider our understanding of ourselves, the world, or things like crystal structures or Go.[21]

On the basis of these two distinctions, there will be a continuum between each of the four characteristics of surprises: just as the transition from 'mere' prediction failure to model disruption will be blurry, so is the distinction between a surprise that is co-produced v. a surprise that is solely made possible by 'downstream interpretation'. What matters, for our purposes, is that only one type of surprise is meaningfully related to both Lovelace's originality insight and Turing's inquiry into the plausibility of machines ever 'thinking': co-produced surprises that prompt some model change ('4' in Fig. 1 below). Trivial surprises ('1' and '3'), which merely highlight a failed calculation or prediction, are neither here nor there when it comes to a machine's purported 'thinking abilities'. We have had mechanical devices capable of highlighting mistakes in our calculations since Antiquity. The same goes for surprises that are made possible solely by the recipient's interpretation ('1' and '2'): Nabokov's 'white parallelogram of sky being unloaded from the van' turns out to be a dresser (with mirror) (Nabokov, 2001). Neither the dresser nor the person carrying it had any originality ambitions—or thinking pretensions (the dresser could have been unloaded by a mechanical robot)[22] yet that object triggers a surprise that sits somewhere between 'mere' prediction failures and model changing encounters.[23]

[20] Skorin-Kapov (2015) distinguishes between what she calls 'irreducible' and 'banal' surprise, the latter denoting an event that was initially and 'accidentally' experienced as surprising. This contrasts with my proposed 'anticipation failure' criterion, which leads to surprises that can but need not be accidental.

[21] Margaret Boden (2016, pp.70-71) draws a related distinction when she contrasts the improbable, 'statistical surprise' generated by 'combinational creativity' with the 'deeply surprising ideas' generated by what she calls 'exploratory creativity'. She highlights that these deeply surprising ideas are 'often initially unintelligible, for they can't be fully understood in terms of the previously accepted way of thinking'.

[22] Conversely, I may fail to notice (let alone be surprised by) a creative act that is deemed original by many others.

[23] 'The unloaded mirror shares with Novalis's romantic aesthetic the central goal of recovering what we already know, rather than discovering, the goal of bringing into consciousness, the conditions of daily life, the repressed, the habitual or the forgotten script in which events take place. None of these cases leads to learning, and they are not, in Descartes's sense, an attention to something new.' (Fisher, 1998, p. 28).

That AlphaGo was capable of generating a move that triggered some model change for GO experts is a feat that reflects that system's degree of operational autonomy, itself resulting from a sophisticated learning process. Among the 10^170 possible moves, 'move 37' was a generative move, in that it prompted a fundamental shift in GO experts' understanding of the game. Yet, that move did not 'originate' anything. The Go experts did. They could have seen AlphaGo's unprecedented move as just one more 'new' move and leave it at that (in that case it would have been located in '1' in Fig. 1 above). Instead they saw that this move called into question commonly held strategic assumptions. They were surprised. This surprise was made possible by the experts' own interpretation, which not only required a sophisticated understanding of the assumptions that were questioned by that move. It also required a readiness to 'see' that move as surprising.

That readiness cannot be taken for granted, and points at two contrasting aspects of the relationship between learning and surprise. In the absence of—or prior to—learning processes, 'everything [is] unexpected because there [is] not yet in place that fundamental feature of mature experience, the idea of an ordinary world'.[24] Yet once learning has shaped our expectations, it can end up preventing us from seeing the 'truly unexpected':

> Since wonder declines with age, as Descartes pointed out, it is a basic experience of youth, but only of youth far enough along into the familiarity of the ordinary world to be able to see against this background the truly unexpected, the truly beautiful. (Fisher, 1998, p. 19)

Interestingly, a compromised ability to 'see' something as surprising—let alone experience wonder[25]—is one of the key obstacles to the successful deployment of systems that are designed to learn to navigate unknown or changing environments. In order to preserve such systems' 'ability to be surprised', computer scientists have sought to quantify the notion of surprise. These quantification endeavours tend to overlook two important limitations. First, they become inept when it comes to co-produced, non-trivial surprises. The second, unacknowledged limitation is related to the first: the concept of surprise is not value-neutral. While certain types of surprises (such as the '1', prediction failure type in Fig. 1) tend to be associated with a negative valence, other types (such as the non-trivial surprises in '3' and '4') tend to be positive events whose value need not be solely epistemic. A world without the co-produced, non-trivial surprises encompassed in '4' would also be poorer aesthetically and ethically speaking (Morton, 2014). The next section problematises current endeavours to quantify surprises.

[24] 'Our first experiences, our first sight of the sun, snow, fire, the stars at night took place in infancy or early childhood at a time when everything was unexpected because there was not yet in place that fundamental feature of mature experience, the idea of an ordinary world' (Fisher, 1998, p. 19).

[25] Because wonder, as Fisher puts it, 'does not depend on awakening and then surprising expectation, but on the complete absence of expectation', it is qualitatively distinct from the experience of surprise, whether trivial or not.

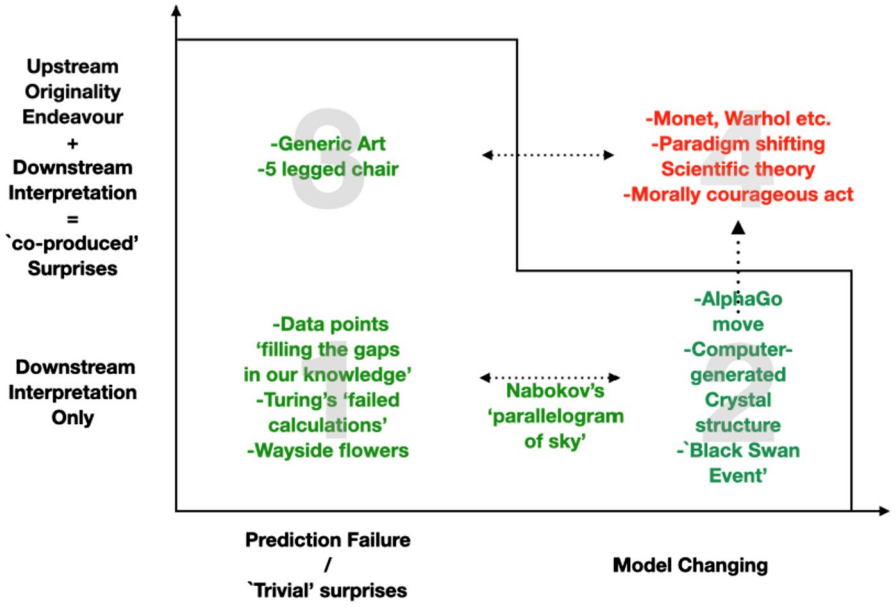

Fig. 1 Types of surprises

4 Optimising a System's Learning Capacity Through Artificial Curiosity or 'Surprise' Constraints

The difficulty inherent in preserving a system's ability to recognise the extent to which some input challenges a given model is commonly associated with Bayesian learning methods: as the number of data samples increases, model uncertainty can reach close to zero, in turn compromising the learning capacity of the system. The latter indeed requires a balance to be found between the plasticity necessary to being able to draw upon new knowledge-generating experiences and the stability without which learned memories get forgotten. This insight is far from new. Yet, its re-discovery within the field of Machine Learning is invigorating surprise-focused research.[26]

This includes renewed attempts to quantify surprises. A Bayesian approach focuses on the difference between prior and posterior beliefs about model parameters (the greater that difference, the more 'surprising' the data is), whereas Shannon's version takes a model as given and aims to capture the inherent unexpectedness of some data (the latter has also been called 'surprisal') (Tribus, 1961, pp. 64–66). The two approaches are complementary and are increasingly relied on to

[26] While there are clear (and interesting) links between the two, one has to distinguish between surprise-related literature that is focused on improving the learning performance of the system (discussed above) and research that is concerned with maintaining the interest of the system's end-users by introducing 'serendipitous' outputs, as in recommender systems. For a survey, see Kotkov et al. (2016).

devise mathematical tools to optimise the extent to which a learning algorithm suitably learns (and keeps learning) from available—or emerging—data points (Faraji, 2016; Faraji et al., 2016). These tools are mostly designed to ensure that 'a small model uncertainty remains even after a long stationary period', thus ensuring improved learning performance in changing, dynamic environments (Faraji, 2016, p. 39).[27]

Because it is designed to rely on the information that is gathered at each stage to adapt or fine-tune its parameters, a learning algorithm is by necessity only as good as the data it has been fed. If it has to function in a dynamic environment, or at least one that is not as stable as anticipated, the learning algorithm is likely to produce sub-optimal results. The problem is that the algorithm itself is unlikely to be able to grasp the instability of the environment within which it is made to function, unless extra, 'artificial curiosity' constraints are introduced Ngo et al., 2012; Storck et al., 1995). The latter are meant to prompt the system to actively extend the range of instances over which it has data ('1' in Fig. 1). By looking for uncommon, 'black swan events'[28] that might demand some model alteration ('2' in Fig. 1), these curiosity constraints are meant to counter-balance the Bayesian tendency towards near-zero model uncertainty.

This concern to maximise information gain typically underlies the design of systems designed to navigate dynamic environments. The latter may improve their learning performance if they 'plan to be surprised', according to Sun et al. (2011). In that case, the type of surprise involved will congregate towards '1' (in Fig. 1 above), including also some model changing surprises of type '2'. When a particular application demands that uncertainty be minimised, however, one may devise what Faraji et al. (2016) call a 'surprise-minimisation rule', to refer to a 'learning strategy which modifies the internal model of the external world such that the unexpected observation becomes less surprising if it happens again in the near future'. This is something humans tend to do quite a lot: when encountering a highly unexpected event, we tend to give a lot of weight to the latter, often discounting past knowledge (sometimes overly so). This parallel is not lost on Faraji et al. (2016), whose ambitions go beyond the design of mathematical tools to optimise a system's learning performance. They believe that their 'surprise-modulated belief update algorithm […] can be used for modelling how humans and animals learn in changing environments'.

At this point, those philosophers who had merely raised their eyebrows at the idea that one may sensibly seek to quantify surprises may shift to genuine alarm. It is one thing for a rich concept such as surprise to be used narrowly by computer scientists—mostly ignoring type '4' in Fig. 1—for their own purpose. An endeavour to improve the learning performance of an algorithm designed to navigate some maze, or to play some game, need not—at this stage at least—be overly concerned

[27] This remaining uncertainty ensures that an organism can still detect a change even after having spent an extensive amount of time in a given environment (Faraji, 2016, p. 39).

[28] The data-mining literature that focuses on the detection of anomalies within noisy data-sets (Eskin, 2000) proceeds from a different starting point (in that the data is given) but the underlying logic is similar.

with the fact that human life could be severely impoverished if the likelihood—or 'amount'—of surprise were to come close to zero (thanks to a very efficient 'belief update algorithm').

Yet it is another thing altogether to overlook the fact that in many domains of human life, error-prediction minimisation only plays a marginal role among the goals that structure our learning to find our way around the world. Among other relevant goals, §2.2. emphasised the drive to originality. The kind of autonomy presupposed by the latter depends on different learning processes from those which enable error-prediction minimisation. Unlike operational autonomy, creative autonomy entails the ability to imagine how things could be different (including the norms that structure one's environment). To be capable of originality, this creative autonomy needs to be supplemented with the interpretive capabilities outlined in §2.3.[29] Today's systems (just like Turing's) are still far from achieving such a degree of interpretive sophistication.

5 Conclusion: Why Lovelace's 'Originality Insight' Is Still Valid Today—Just as Much as Turing's 'surprise' Translation

The above heading should not be confused with some misplaced diplomatic effort: while Lovelace's originality insight remains correct, the way Turing translated that insight is both odd and fascinating in equal measure.

Why odd? Lady Lovelace never refers to the concept of surprise. Having some pretension to 'originate' something—unlike the Analytical Engine—is neither necessary nor sufficient to being able to surprise someone: a rainbow may surprise me. Conversely, I may not be surprised at all by the fact that my friend has just come up with an original theorem. So, why does Turing translate Lady Lovelace's point about the lack of pretension to originate anything into a claim about the ability to surprise? Was he influenced by his 1943's discussions with Shannon? Perhaps. But 'Shannon surprises' are not even remotely connected to Lovelace's point about the pretension to 'originate' something (and at any rate it is unclear whether Shannon was even considering using the term 'surprise' in the 1940s).

Why fascinating? Because what is 'lost in translation' in Turing's odd interpretation of Lovelace happens to be a key hurdle in the design of systems capable of navigating the world we live in. Unlike games, the norms that structure our quotidian lives are necessarily dynamic. As our needs and aspirations change, so do the social, cultural and moral norms that govern our interactions with our world and its inhabitants. As we marvel at the ingenuity—and unexpectedness—of AlphaGo's 'move 37', or the Rembrandt-like appearance of assembled pixels, it is easy to lose sight of the limits inherent in different types of learning processes, and the kinds of autonomy they facilitate (Nudd, 2016).

[29] Without an understanding of the 'the unoriginal structures that inform originality', creative autonomy can remain 'sterile', out of reach of second-person perspectives.

To 'do as children do', the systems we build would not only need the kind of operational autonomy brilliantly displayed by AlphaGo, or the apparent[30] creative autonomy[31] encapsulated in systems meant to generate art (Kosoy et al., 2020, p. 4). They would also need 'hermeneutic autonomy': an ability to interpret and re-appropriate the fabric of socio-cultural expectations that can be (and regularly is) transformed through creative intervention. In short, they would need a grasp of the significance of what §2.3 discussed as the 'second-person dimension' of originality. Such a grasp should not be confused with mastery: for children and adults alike, the conditions that preside over the receptivity and understanding of our fellow humans are largely opaque. The extent to which they remain so depends in part on our varying hermeneutic skills, as well as a willingness to engage with (and sometimes test) this receptivity through originality endeavours. When the latter fail to generate any reaction (surprise or otherwise), the originality effort was not necessarily in vain. Not only does the absence of reaction bring some insight into our fellow humans' receptivity, that originality endeavour is also of value in and of itself, given its role in the process of identity formation and collective transformation.

The latter considerations bring to the fore the question first raised in §2.2.: if what drives originality as an effort is the need to challenge the web of socio-cultural expectations that shapes us from birth, one may query its relevance as a criterion for machine intelligence. If, unlike us, machines are able to control what they let themselves be 'shaped' by, the most helpful translation of Lovelace's originality insight turns out to be an upside-down version of Turing's surprise question: can computers be surprised *by us* in a non-trivial, 'co-produced' way? Aside from its doubtful desirability, a positive answer would require a leap in our digital machinery's interpretive capabilities.[32] Such a leap is unlikely to be adequately captured in any game-based test, whether Turing's or otherwise.

[30] When a system that has been 'fed' large numbers of Rembrandts produces an assemblage of pixels that looks like it could have been painted by Rembrandt, are we faced with art? Or are we 'merely' faced with some artefact that can be used by human artists as a starting point in their exploration of different ways of re-interpreting some past heritage to originate something new? Du Sautoy (2019) rightly resists the temptation to adopt an entirely 'downstream' definition of art, according to which artistic status solely depends on emotional responses. Yet, to insist that art has to be about 'exploring what it means to be a conscious emotional human being' probably goes too far the other way. A less anthropocentric criterion considers whether some output is the product of an effort of interpretation of some shared past (or present), as discussed above. None of the current, 'AI generated art' is. It might be beautiful, intriguing or a useful source of inspiration to human artists. But to count as art, the algorithm producing it would have to be able to claim that it has sought to interpret its surrounding world, inhabited as it is by both humans and machines.

[31] The term 'creative autonomy' is favoured over 'creativity' in part because the question of whether 'learning machines' may ever *experience* the processes that lead to the generation of an original output in such a way as to warrant the use of the term 'creativity' is beyond the scope of this paper: see Nanay (2014), who emphasises the superiority of 'experiential' accounts of creativity over their functional/computational counterparts.

[32] Along this line, Gonzalez and Haselager (2005) emphasise that what they 'take to be great cause for caution in speaking about surprise in artificial systems (computer programs, robots) is that they appear to be extremely limited in both the type of disturbances they can notice and the experiential effects these disturbances can have on them'.

Acknowledgements I am grateful to Chris Baber, Rudolf Bernet, Simon Blackburn, Mireille Hildebrandt, Neil Lawrence, Bence Nanay, Jurgen Van Gael, Michael Veale and Alan Wilson for their helpful comments. I also benefited from excellent feedback from the University of Cambridge CFI seminar, the Birmingham computer science seminar and a presentation at the Bellagio AI Conference.

Funding The research leading to this work was funded by the Leverhulme Trust, as well as a Mozilla research Fellowship.

References

Abramson, D. (2008). Turing's responses to two objections. *Minds and Machines, 18*, 147–167. https://doi.org/10.1007/s11023-008-9094-6.

Arendt, H. (1971). Thinking and moral considerations: A lecture. *Social Research*, 417–446.

Arendt, H. (1989). *Lectures on Kant's political philosophy*. University of Chicago Press.

Arendt, H. (1998). *The human condition*. The University of Chicago Press.

Arendt, H. (2007). *The promise of politics*. Schocken.

Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence, 103*(1), 347–356. https://doi.org/10.1016/S0004-3702(98)00055-1.

Boden, M. A. (2010). *Creativity and art: Three roads to surprise*. Oxford University Press.

Boden, M. A. (2016). *AI, its nature and future*. OUP.

Bringsjord, S., Bello, P., & Ferrucci, D. (2001). Creativity, the Turing test, and the (better) Lovelace test. *Minds and Machines, 11*, 3–27.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Deringer, V. L., Proserpio, D. M., Csányi, G., & Pickard, C. J. (2018). Data-driven learning and prediction of inorganic crystal structures. *Faraday Discussions, 211*, 45–59. https://doi.org/10.1039/C8FD00034D.

Du Sautoy, M. (2019). *The creativity Code: How AI is learning to write, paint and think*. Harper Collins.

Eskin, E. (2000). *Anomaly detection over noisy data using learned probability distributions*. Paper presented at the Proceedings of the International Conference on Machine Learning.

Faraji, M. (2016). *Learning with surprise: Theory and applications.* École Polytechnique Fédérale de Lausanne

Faraji, M. J., Preuschoff, K., & Gerstner, W. (2016). *Balancing new against old information: The role of surprise*.

Fisher, P. (1998). *Wonder, the rainbow, and the aesthetics of rare experiences*. Harvard University Press.

Gaut, B. (2003). Creativity and imagination. In B. Gaut & P. Livingston (Eds.), *The creation of art* (pp. 148–173). Cambridge University Press.

Gaut, B. (2010). The philosophy of creativity. *Philosophy Compass, 5*(12), 1034–1046.

Gonzalez, M. E. Q., & Haselager, W. F. G. (2005). Creativity: Surprise and abductive reasoning. *Semiotica, 153*(153), 325–341. https://doi.org/10.1515/semi.2005.2005.153-1-4.325.

Hinchman, L. P., & Hinchman, S. K. (1984). In Heidegger's shadow: Hannah Arendt's phenomenological humanism. *The Review of Politics, 46*(2), 183–211.

Jarvie, I. C. (1981). The rationality of creativity. In D. Dutton & M. Krauz (Eds.), *The concept of creativity in science and art* (pp. 109–128). Springer.

Kosoy, E., Collins, J., Chan, D. M., Hamrick, J. B., Huang, S., Gopnik, A., & Canny, J. (2020). *Exploring exploration: Comparing children with RL agents in unified environments*. Paper presented at the Bridging AI and Cognitive Science (ICLR 2020).

Kotkov, D., Wang, S., & Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems, 111*, 180–192. https://doi.org/10.1016/j.knosys.2016.08.014.

Lovelace, A. A. (1843). Notes by A.A.L. [August Ada Lovelace]. *Taylor's Scientific Memoirs, III*, 666–731.

Morton, A. (2014). Surprise. In *Emotion and Value*. Oxford University Press.

Nabokov, V. (2001). *The Gift*. Penguin.

Nanay, B. (2014). An experiential account of creativity. In E. S. Paul & S. B. Kauffman (Eds.), *The Philosophy of Creativity: New essays*. Oxford University Press.

Ngo, H., Luciw, M., Foerster, A., & Schmidhuber, J. (2012). *Learning skills from play: Artificial curiosity on a katana robot arm.* Paper presented at the Proceedings of the 2012 International Joint Conference of Neural Networks.

Novitz, D. (1999). Creativity and constraint. *Australasian Journal of Philosophy, 77*(1), 67–82.

Nudd, T. (2016). Inside "The next Rembrandt": How JWT got a computer to paint like the old master. *Adweek*. Retrieved from https://www.adweek.com/brand-marketing/inside-next-rembrandt-how-jwt-got-computer-paint-old-master-172257/.

Ricoeur, P. (1975). *Lectures on imagination*. Unpublished, recordings held at University of Chicago.

Silver, D., Huang, A., Maddison, C. J., Guez, A., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. https://doi.org/10.1038/nature16961.

Skorin-Kapov, J. (2015). *The aesthetics of desire and surprise: Phenomenology and speculation*. Lexington Books.

Srinivasan, A. (2019). Genealogy, epistemology and worldmaking. *Proceedings of the Aristotelian Society, CXIX*.

Storck, J., Hochreiter, S., & Schmidhuber, J. (1995). Reinforcement driven information acquisition in non-deterministic environments. *Proceedings of the International Conference on Artificial Neural Networks, 2*, 159–164.

Sun, Y., Gomez, F., & Schmidhuber, J. (2011). *Planning to be surprised: Optimal Bayesian exploration in dynamic environments*. Paper presented at the International Conference on Artificial General Intelligence. http://people.idsia.ch/~juergen/agi2011sun.pdf.

Taylor, G. (2006). Ricoeur's philosophy of imagination. *Journal of French Philosophy, 16*, 93.

Tribus, M. (1961). *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering applications*. D. Van Nostrand.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Turing, A. M. (2004). Can digital computers think? In B. J. Copeland (Ed.), *The Essential Turing*. Oxford University Press.

Turner, M. (1991). *Reading mlnds: The study of English in the age of cognitive science*. Princeton University Press.

Williams, B. (2000). Philosophy as a humanistic discipline. *Philosophy, 75*(04), 477–496.