Steve Benford, Adrian Hazzard, Craig Vear, Helena Webb, Alan Chamberlain, Chris Greenhalgh, Richard Ramchurn, and Joe Marshall. 2023. Five Provocations for a More Creative TAS. In First International Symposium on Trustworthy Autonomous Systems (TAS '23), July 11–12, 2023, Edinburgh, United Kingdom. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3597512.3599709

Five Provocations for a More Creative TAS

Steve Benford

University of Nottingham, steve.benford@nottingham.ac.uk

Adrian Hazzard

University of Nottingham, adrian.hazzard@nottingham.ac.uk

Craig Vear

University of Nottingham, craig.vear@nottingham.ac.uk

Helena Webb

University of Nottingham, helena.webb@nottingham.ac.uk

Alan Chamberlain

University of Nottingham, alan.chamberlain@nottingham.ac.uk

Chris Greenhalgh

University of Nottingham, chris.greenhalgh@nottingham.ac.uk

Richard Ramchurn

University of Nottingham, richard.ramchurn@nottingham.ac.uk

Joe Marshall

University of Nottingham, joe.marshall@nottingham.ac.uk

Conventional wisdom has it that trustworthy autonomous systems (AS) should be explainable, dependable, controllable and safe tools for humans to use. Reflecting on a portfolio of artistic applications of TAS leads us adopt an alternative stance and to propose five provocative challenges for AS: that they should look beyond failure to improvisation, beyond explainability to interpretation, beyond control to surrender, beyond caution to playfulness and beyond being tools to becoming co-creators. We reflect on how these challenges imply new considerations of trustworthiness in terms of artistic competence, sincerity and steadfastness, and of responsible innovation in terms of responsible irresponsibility that empowers humans to playfully explore the human-like qualities and boundaries of the technologies.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI)~HCI theory, concepts and models•Applied computing~Arts and humanities~Media arts • Human-centered computing~Collaborative and social computing~Collaborative and social computing theory, concepts and paradigms

Additional Keywords and Phrases: Autonomous systems, art, trust, responsibility

ACM Reference Format:
First Author's Name, Initials, and Last Name, Second Author's Name, Initials, and Last Name, and Third Author's Name, Initials, and Last Name. 2018. The Title of the Paper: ACM Conference Proceedings Manuscript Submission Template: This is the subtitle of the paper, this document both explains and embodies the submission format for authors using Word. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY. ACM, New York, NY, USA, 10 pages. NOTE: This block will be automatically generated when manuscripts are processed after acceptance.

1 INTRODUCTION

Autonomous systems, and AI more generally, are being used to make art, from generative visual art to music composition, to dancing robots, and much more besides. In turn, such artistic explorations can offer fresh perspectives on the human-experience of autonomous systems and related questions of trustworthiness and responsibility. With this in mind, the Trustworthy Autonomous Systems (TAS) Hub is already supporting a series of artistic explorations through its Creative Programme of artist-led projects and residencies.

Whereas applications in fields such as transportation and healthcare naturally stress the importance of explainability, dependability and controllability and view TAS as tools to be applied with a degree of caution, artistic applications emphasise interpretation, improvisation, surrendering control, playfulness, and the creative possibilities of viewing TAS as co-creators alongside humans.

We reflect on a number of artistic case studies from the TAS Hub and beyond involving autonomous systems to draw out five key provocations for more creative trustworthy autonomous systems (TAS). Our aim is to provoke the TAS community to question what appear to be almost obvious – perhaps even taken for granted – assumptions underlying TAS to instead consider alternative possibilities in which TAS actively embrace ambiguity, uncertainty, and playfulness as part of a more creative, and we argue, human-like approach to the world. Throughout, we illustrate our challenge by reflecting on previous and ongoing examples of creative human engagement with a variety of AS spanning robots, affective computing and AI music composition and performance.

2 BEYOND FAILURE TO IMPROVISATION

Error and failure are traditionally seen as problems to be avoided and removed in the design of dependable computer systems [1]. In contrast, error, failure, and fragility can be vital elements of artistic processes inspiring creativity and demanding improvisation. The AHRC-funded Error Network, for example, explored how errors range from perilous mistakes to creative impulses and how 'noise' and glitches in digital code afford new possibilities for creative process in art, dance, music and games [2].

Our first example, *Climb!*, a musical duet for a human concert pianist and a robotic piano (Yamaha Disklavier) by composer and pianist Maria Kallionpää [3]. The piece took the form of a game in which the human partner had to correctly perform musical challenges – the execution of pre-scored musical fragments or 'codes' – in order to choose routes through a non-linear score (see fig. 1). These codes also triggered other performance media such as the piano to autonomously play its parts, including physically moving its keys in duet with the human performer. The piece set up a contest between a highly skilled human and an apparently skilled instrument capable of playing very quickly and accurately, reaching combinations of notes that a human would find impossible, and doing this without tiring.

Figure 1. Performing Climb!

Studies of three professional concert pianists performing *Climb!* revealed failure to be essential to the aesthetic of the work and also to be a complicated layered phenomenon [3]. Deliberately introducing errors into the performance became a first-class mechanism for the pianists, for example branching points were written into the score in the form of the 'challenge codes', deliberately difficult musical fragments that if performed correctly or incorrectly (from the perspective of the system) determined how the piece progressed – consequently, it was important for errors to occur so the performer could vary their route through the score. Without them, a performance would be dull and predictable. Conversely, failure to successfully trigger other codes would result in the piano failing to play its autonomous parts, which then required rapid improvisation by the human pianist to keep the performance on track. Even if the piece proceeded smoothly, it might be musically unexpressive, a different kind of error that required the human to add a layer of expression to the system's somewhat mechanical performance. Extreme divergence from the score, however musical, might result in failure to perform a recognisable version of the work (i.e. to sufficiently adhere to the score). Finally, there was one occasion where the system failed catastrophically, requiring the performer to apologise to the audience for failing to deliver any kind of musical performance at all [3].

This study of how humans performed *Climb!* also yielded insights into their strategies for creatively engaging autonomous systems, with them variously trying to 'tame' the piano, 'game' it by exploiting its weaknesses, 'ride' it in which they improvised around whatever it did, and 'become' it when they strictly conformed with its needs as if a component in the system [3].

Might other less creative AS benefit from similar strategies? Conversational agents are one area where failure can be a multi-level, improvisational and creative matter. Ethnography and conversation analysis revealed how Alexa becomes embedded into and held accountable as part of everyday family dinner time conversations [4]. Puns, irony, sarcasm and other forms of humorous wordplay deliberately invoke failures of language at one level to deliver humour when people reflect at a meta-level, something that is currently challenging for AI to achieve. Turning to autonomous vehicles, anecdotally, one of the authors experienced a self-parking car managing to parallel park itself in a very tight space, but taking so much time to do so as it repeatedly moved back and forth that a queue

of traffic built up behind – an example of success at one level (the technical task of parking) but failure at another (achieving this in a socially acceptable way). More generally, plans often fail to survive their first encounter with reality, being treated instead as resources for negotiating and improvising action rather than as procedures to be strictly followed [5]. The implication is clear: AS should be designed to embrace or even deliberately invoke a variety of aesthetic failures.

3 BEYOND CONTROL TO SURRENDER

It is a longstanding tenet of human-computer interaction (indeed one of Scheiderman's 'eight golden rules') that the locus of control should lie with the human when interacting with computer systems [6, 7]. However, this is challenged by the emergence of autonomous systems which, by definition, wrest some degree of control from humans, raising profound questions of trust. In artistic applications humans may deliberately choose to surrender a degree of control to the system as a creative strategy.

In music, the composer John Cage spent much of his creative life moving away from making and imprisoning sound into a fixed composition to creating more open works. Cage's approach was inspired by his devotion to Zen Buddhism and other Eastern philosophies, and eventually led him to explore environmental and unintended sounds as materials for music (performance and composition). Cage's milieu has had a profound impact on the role of the composer and the co-creative potential of the performer-as-author in music–making. For many decades, musicians have adopted this philosophical approach, and this is having an impact on those who create with AS. For example, Arne Eigenfeldt's *MuseBots* [8] are autonomous systems that perform together or with human improvisers. They have been embedded with certain traits of behaviour and are designed to collaborate together. The music that emerges is accepted by Eigenfeldt and co-creator Ollie Bown, they do not wish to control or edit it in any way. In this sense they trust the core design of the bots, and their reason-d'etre; nothing more is required of them.

Vear's *Embodied Musicking Robots* (EMR) is another music-bot project also built on Cage's philosophy of acceptance [9]. These AS were designed to behave in real-time and in the embodied domain of music-making. They are embedded with simple rules (inspired by Rodney Brook's early works) and are trusted to behave as a coping machine within musicking. Vear [9] generated a set of principles that guided the design of the AS which in turn controls their behaviour:

- EMR must cope in an appropriate musical manner, and in a timely fashion, with the dynamic shifts inside the musicking world.
- EMR should be robust to the dynamic environment of musicking, it should not fail to minor changes in the properties of the flow of musicking and should behave appropriately to its ongoing perception of the flow.
- EMR should maintain multiple goals, changing as required and adapting to its world by capitalising on creative opportunity.
- EMR should do something in the world of musicking, 'it should have some purpose in being' [10].

The experience of surrendering control and making oneself vulnerable is a powerful one that is harnessed by many artistic and cultural experiences that involve journeys through suspense and fear, from horror films to rollercoasters. It is also one that directly invokes questions of trust. In this vein, several artists have explored how it feels to surrender control to autonomous systems.

The Broncomatic was an interactive ride created as part of the Thrill Laboratory project with artist Brendan Walker [11]. A rider would mount a robotic rodeo bull that would try to throw them off, gradually increasing its

bucking motion (see fig. 2). A chest strap sensor estimated their breathing; the more they breathed the more the ride twisted, but also the more points they scored.

Figure 2: Riding The Broncomatic

Studies of *The Broncomatic* revealed how riders simultaneously battled for control of the robot and their own bodies (as breathing is partially an autonomic system) [11]. A highlight of the experience for some was when they were sitting on top of the robot, trying to control it by holding their breath, but becoming aware that their breath would soon run out at which point the ride would buck even more as they then breathed deeply and uncontrollably [11], thus handing over or surrendering control to the AS.

The example of *#Scanners* – further illustrates surrender of control [12, 13]. This project saw the creation of two brain-controlled films, where real-time measurements of a single viewer's physiology dynamically adapted the film (see fig. 3). In *The Disadvantages of Time Travel*, estimates of attention, meditation and blinking derived from EEG triggered switching between alternative layers of film representing the protagonist's external and internal worldviews [12]. EEG measurements of average levels of attention at the end of each scene were used to decide which characters' viewpoints would feature in the next scene [13].

Figure 3. Watching brain-controlled films in #Scanners

Studies of how viewers experienced *The Disadvantages of Time Travel* highlighted tipping back and forth between controlling the film and being controlled by it, sometimes trying to consciously control the film, at others forgetting about (surrendering) control as they became absorbed in the story, and at yet others being reminded of the uncontrollable nature of their own inner thoughts [12].

Reflection on these various experiences revealed how the human experience of control over autonomous systems can be considered as traversing a space comprised of two dimensions, the first expressing the degree to which humans surrender control to the system or otherwise, and the second expressing their shifting awareness of being in control or otherwise [14]. The resulting interactions range from humans voluntary exerting control and being aware of it, to flow states [15] in which they are in control but without overt awareness, to states in which they are aware of losing control (paused on top of *The Broncomatic!*), to losing both control and self-awareness.

Again, we invite the reader to consider how these ideas apply beyond artistic applications. How might people enjoy the experience of surrendering control and to what extent will they even be aware of it. Moreover, as embodied autonomous systems such as robots increasingly make contact with – and therefore stimulate – our human bodies, to what extent must we also contest control with our own physical and partially autonomic selves.

4 BEYOND EXPLAINABLE TO INTERPRETABLE

Conventional wisdom has it that current AS and AI more generally need to be more explainable [16, 17]; that is, AI systems should include mechanisms to ensure their predictions and decisions can be understood by humans, and that techniques should be used to ensure humans can interpret the reasoning behind system outputs accurately. However, AI art offers an alternative vision in which AI is designed to be open to multiple interpretations by humans. Frequently, art does not explain itself. Nor is it explained by its creators. Rather it demands that audiences, critics and scholars form their own interpretations. HCI research has previously discussed how art deliberately invokes ambiguities of information, context, and relationship to become open to multiple interpretations, proposing that these strategies might be applied to interaction design [18, 19]. Might such strategies also apply to AI?

Our next case study, *Sensitive Pictures* was a project led by games and media design company Next Game to explore visitors' emotional engagement with Edward Munch's famously emotional paintings at the Munch Museum

in Oslo [14]. Visitors were invited to select and view Munch paintings while listening to audio stories as short emotional provocations before self-reporting their emotional responses. During a subsequent scripted phone conversation with a fictional portrayal of Munch, computer vision technology attempted to detect the emotions they might be feeling from their facial expressions (see fig. 4). At the end, they were presented with postcards summarizing the emotion data they had generated during the visit. *Sensitive Pictures* does not explain itself to visitors but provides a distorted mirror to encourage people to interpret their own emotional journeys through the museum. This implies a shift away from the idea of affective computing in which computers objectively 'measure' and adapt to human emotions towards one of affective interaction in which they provoke emotions and emotional reflection [20]. In a similar vein, *#Scanners* [12] provoked people to reflect on their own thought processes as viewers.

Figure 4: A phone conversation with AI Munch uses computer vision to try to detect human emotions.

Continuing the discussion with the Vear's *Embodied Musicking Robots* (EMR) example from above [9], it is almost impossible to clearly articulate the cognitive and embodied processes that happen inside the creative acts of music-making (to this end Vear has created a podcast series where he attempts to discuss this phenomenon with a range of experienced musicians). To circumvent this impossible task of explaining what his AS are doing, he prefers to offer behavioural principles of how they should approach the co-creative relationships inside music. These principles manifest in behaviour and affectual relationships between the improvising humans and AS. Aside from circumventing the difficulty of explaining what an AS needs to do, these principles focus on a core tenet in musicking that 'the act of musicking establishes in the place where it is happening a set of relationships, and it is in those relationships that the meaning of the act lies' [21]. Meaning in music (or the what-you-mean-to-me) is naturally a subjective interpretation, it therefore makes sense that the behaviours of the EMR AS focus on stimulating such relationships.

Climb! [22] provides an audience web app to enable audience members to follow the performer's journey in real-time through this non-linear work. They receive notifications of impending musical challenges and post-challenge outcomes of success or failure. Thus, the audience are aware of the pianist's playful negotiation through potential failures and contesting of control, but nonetheless trust a 'performance' will be presented. A study of the audience experience of *Climb!* revealed how people were able to make gradually unfolding interpretations of the work over multiple repeat performances [22].

More generally, the rapidly growing movement of generative art in which machine learning is used to generate artistic images, sometimes with the recognizable style of human artists, is introducing new ambiguities of information (overfitting and underfitting machine learning models), context (appropriating training data for new purposes) and relationship (raising questions of whether AI can make and own art) throughout a complex AI pipeline, all of which serves to make AI less explainable and more open to interpretation. There is currently much public debate over the ambiguous nature of such work, including over its provenance and whether it is art at all. Broadening the discussion further, considerations of responsible TAS are rightly concerned with the important question of bias [23, 24, 25]. Adopting our interpretative stance suggests a possible alternative approach to bias in which AI is overtly positioned as being interestingly opinionated (as human often are), with its biases revealed, perhaps amplified, open to challenge, and again, acting as a distorted mirror to the humans who trained it.

5 BEYOND CAUTIOUS TO PLAYFUL

We are understandably cautious about the extent and nature of direct human contact with AS, especially with robots that might cause people physical harm. Much effort has been invested in enabling autonomous vehicles, drones, and telepresence robots to avoid collisions and to mitigate them when they do occur [26].

In contrast, *The Broncomatic!* [11] invites direct physical engagement with a robotic AS, one that clearly instigates collisions as the rider is thrown through the air to land on the cushioned floor below. It also establishes a playful relationship with the technology in the tradition of rollercoasters and other amusement rides, the human is invited to experience an apparent degree of risk in the interests of thrill. This is, however, a carefully managed risk in the presence of many safety features including an inflatable cushioned arena, wearing safety equipment, and a vigilant human operator with a shut-down control.

Our other examples also establish playful relationships with the technology that involve negotiating different kinds of risk. *Climb!* [3, 22] is overtly framed as a game that pits a trained concert pianist against a robotic piano. There is tension in the performance arising from the emotional (rather than physical) risk of embarrassment that that the human pianist will be unable to play as well as the mechanized instrument or will fail to meet the musical challenges they have been set in the score.

Vear's *Embodied Musicking Robots* [9] were guided by the following principles that sought to establish an appropriately playful robot with the human composer and performer:

- The robot was not an extension of the musician; but should extend their creativity.
- The robot should not be an obedient dog or responsive insect jumping at my commands or impetus, but a playful other.
- It should not operate as a simulation of play, but as a stimulation of the human's creativity.
- It is not a tool to enhance the human's creativity, but a being with presence in the world that they believe to be co-creating with them.

- It should prioritise emergence, surprise, mischief, not expectation.

Reflecting more generally on these various examples, play is a vital aspect of being human and often involves negotiating a degree of risk and discomfort, physical or emotional [27]. Play allows us to test out our relationship with the world – children learn to fall and collide through rough and tumble [28, 29], skills that serve them well in later life when they encounter more serious collisions. It also allows us to test and improve our human aesthetic skills – many adults continue to play contact sports in later life. Play, and the element of discomfort it may potentially involve, can deliver entertainment, but also personal enlightenment and social bonding, leading to the ides of deliberately designing 'uncomfortable interactions' as part of interactive experiences [30].

We argue that as embodied AS increasingly enter the busy and relatively uncontrolled controlled spaces of our everyday lives, most notably our homes, we will need to playfully engage them in order to test their limits, learn how to engage, and also to assert ourselves as humans. AS may help us play better, for example enabling those with disabilities to play sport or even developing new kinds of contact sport for all. In contrast, risk-adverse robots that continually seek to avoid collision may ultimately unintentionally harm us by boxing us in, disempowering from physically asserting ourselves with our world. Marshall et al have argued for deliberately embracing collisions with robots as a design strategy to avoid them becoming too risk averse [31]. Thus, while we are not arguing for humans playing with autonomous vehicles on motorways, we are arguing for enabling them to play with technologies that enter their living rooms, lawns and other everyday human spaces where people physically assert themselves through play.

6 BEYOND TOOL TO CO-CREATOR

Our final challenge concerns the creative relationship between humans and AS. One possibility is to view AS like conventional computing systems, as a tool to be applied by humans. While this can undoubtedly be true for creative applications – intelligent tools might be introduced into music and video production toolchains for example, our projects suggest other possibilities in which AS becomes a co-creator alongside the human, rather than a tool used by them.

Our final example, *FolkRNN Groover* plays with this co-creative relationship in various ways. The *FolkRNN* project by Bob Sturm and colleagues at KTH Sweden used a public dataset of tens of thousands of transcriptions of traditional folk tunes to train a recurrent neural network to compose its own 'traditional' tunes [32]. The *FolkRNN Groover* project is building on this by exploring how an AI which appears to be embedded into a digitally augmented acoustic guitar – the *Carolan Guitar* [33] – can expressively play these tunes live while a human musician improvises an accompaniment.

FolkRNN Groover establishes familiar co-creative role between human and AS, that of a soloist and accompanist, with the human being the latter [32]. It places the human in a supporting role and shines a spotlight on their capabilities, or otherwise, as an accompanist as they try to bring the soloist's somewhat mechanical performance to life.

A notable aspect of this co-creative relationship is that the AS takes on a distinctive persona. Placing a Bluetooth speaker inside the guitar gives the sense that the AI is embodied as a musical instrument. Moreover, this instrument already has something of a persona – the *Carolan Guitar* has a unique identity, as a kind of digital bard that collects and replays songs and stories from the musicians it encounters [33]. The net effect is to establish the metaphor for human player and audience alike that this is an intelligent instrument with its own autonomous identity. Ongoing

efforts to get the AI to play its tunes with a degree of variation including embellishments and even errors are intended to further reinforce this sense of a musical persona as well as to play with the ideas of error and improvisation discussed above.

A noteworthy aspect of Vear's [9] *Embodied Musicking Robot* series is the design principle of embedding "beliefs" into the code so as to have a distinctive persona and to heighten levels of trustworthiness through continued play. Belief in this sense is a defining of the AS worldview, the umwelt with which it perceives and cognates its world. These beliefs are used to describe an acceptance by the AS that something is true, or to be perceived as true. It is a subjective viewpoint, not a fact. In this sense an AS belief system enables it to trust its world, and therefore the decisions that it makes when engaging with it. An example belief system can be found in his EMRv3 robot [9]:

- Personalised dataset – which grows through each interaction with a musician.
- Movement behaviour – which is limited to a few specific responses (its character).
- Soundworld – its personal theory of music.

Because of these beliefs, and the trust the robot has is its own understanding of how to behave, the musicians who have worked with this AS also grow a sense of trustworthiness, that it will respond and behave in a certain way (it's way) and can then start to grow a sense of trustworthiness with their relationship-making and musicking with these AS. In this sense, trustworthiness grows through multiple confirmations that the AS's behaviours and character are worthy of trust. The human feels that they can trust the bonds and relationships that are generated inside musicking and this can lead to a heightened sense that they can co-create together. Chung et al's [34] artist interviews illustrate how the work of co-creation and collaboration reveals potential points of friction, to which they offer a number of solutions pertinent to our discussion here. Mirroring Vear's [9] insights above, "Rounds of collaboration" [34] were key in developing trust in understanding each other's styles and values, and an awareness of others skill sets helps build trust in their respective skills and qualities, but also in managing expectations. Lee and See [35] highlight the characteristics of calibration and resolution between humans and AS as key in managing expectations. Chung et al [34] also observe communication between artists as a further point of friction, as it often relies on descriptive jargon open to misinterpretation. A mitigating approach is to use 'references or sketches', a mechanism which reflects Vear's [9] call to embed AS with a belief system. Finally, a single artist (actor) in the collaboration should have 'concentration of power' [34] to control decision making, or to dictate or relinquish (surrender) authority to other actors.

This casting of AS as co-creators rather than tools is not to argue that they genuinely possess human-like intelligence – though that is of course has been a concern for AI since its inception as a field. Rather, the idea is that framing the system as a more human-like co-creator might influence the human's creative experience. Do they perform differently, for example being more willing to surrender control, embrace errors and improvisation, and play more freely as discussed earlier, and do they trust the system differently? It also perhaps offers a perspective on the questions of ownership and intellectual property that are at the fore of efforts to employ AI for generative art. How, in practical terms, does the human accompanist introduce and credit the AI in the performance?

7 TRUST AND ART

Having introduced our five challenges we now reflect more widely on how they speak to the two key concerns of trust and responsibility that run throughout the TAS Hub programme.

In their philosophical inquiry "Trust and Sincerity in Art", Nguyen argues for a distinctive treatment of the question of trustworthiness in relation to art [36]. Viewing the matter through an aesthetic rather than a moral lens leads to the idea that an audience might place their trust in some combination of the artist's aesthetic competence to "successfully bring about aesthetically valuable states of affairs" (i.e., deliver 'good' art), aesthetic sincerity to "meet their commitment to act from aesthetic considerations" (i.e., be true to their art and not sell out), and aesthetic steadfastness to "act from their commitment to a specified aesthetic sensibility" (to continue to pursue their artistic aesthetic). They suggest that artistic sincerity is the most important of these, especially for more avant-garde art that seeks to experiment and push boundaries which is consequently difficult to engage with, requiring considerable effort to interpret. While we might feel 'disappointed' when an artist we appreciate fails to deliver good art as a result of a failed experiment, we experience a deeper betrayal of trust when they fail to be sincere, for example 'selling out' for commercial gain or no longer really trying at all. Moreover, whereas conventional moral stances frame trustworthiness as an important social relationship in which one party places their trust in another to actively act on their behalf, this is different for art and artists. We do not expect artists to act in our interests, but rather to sincerely, even narcissistically and obsessively, follow their own, so that we in turn can follow ours and develop our own aesthetic sensibilities. Savery [37] notes that affective trust hangs on the investment of emotional bonds and inter-personal relationships and 'relationships based on affective trust are more resilient to mistakes by either party' [38].

The recent broad interest in text-to-image and text-to-text generative AI systems such as ChatGPT and Dall-E places a further challenge on co-authorship. These systems are trained on large datasets of texts and images taken from the internet. This brings into question whether we trust that the outputs of such systems are (co-)authored by the AI systems themselves with a level of inspiration from the training sets, or whether they in fact are automated plagiarizing systems. For example, if we prompt Dall-E to generate a photo of a badger, it will provide a range of images which look like badgers, with no attribution of which elements of the training dataset influenced the image, and no way to discover how similar image elements are similar to training images. When we collaborate with autonomous artistic systems, there is a risk that they will breach artistic and legal norms relating to copyright. The boundaries between reasonable artistic inspiration and copyright infringement are already hard to define [39]. Introducing a system which gains inspiration from an unknown training set of existing artistic outputs further blurs the boundaries; it is possible that for such systems to be trusted as co-creators we will need to develop new ways of exploring their inspirations.

Given this argument, the question here is to what extent the above artistic uses of TAS are trustworthy with respect to aesthetic competence, sincerity and steadfastness? Some are delivered by professional artists with a longstanding commitment to a given aesthetic and are recognised by a wider artworld, for example Brendan Walker's Thrill Laboratory experiments that date back over many years [30, 40]. Others, most notably *FolkRNN Groover* [32] are more research experiments than trustworthy art. While such experiments may of course be useful to research exploring questions of trustworthy artistic TAS, they may require engaging with established artists whose aesthetic competence, sincerity and steadfastness can be experienced by audiences. This highlights the importance of artist- and practice-led research methods and of the TAS Creative Programme as a mechanism for engaging these.

Further complications arise from the potentially co-creative nature of autonomous systems as discussed earlier. If TAS are themselves co-creators, then how can audiences judge their aesthetic competence, sincerity and steadfastness? AI generated art and music would appear to be taking steps towards – and perhaps is currently being

judged by – levels of artistic competence. They appear to produce works that look and sound something like recognisable art and sometimes even surprise us with what appears to be a creative step. They might also appear to be doggedly steadfast in their pursuit of an aesthetic style. However, it is far harder to trust their artistic sincerity, though we might trust that of the human artists who co-create with them. Abrahams and Kemp [41] in discussing trust and art appreciation remind us of the 'universal definitions of art', that the work is intentionally created and often experienced in a context that further frames this intention. We note that our examples of AS are embedded in established artistic forms that do not fundamentally disrupt the expectations of those who engage with them bring. Rather their implementation is appropriately sensitive to their form. For example, the chamber music work *Climb!* [3] finds a pianist charged with presenting a concert hall performance to a paying audience; whereas the thrill seeker climbs upon the bucking bronco anticipating a bumpy ride and to be thrown to the ground; and the film buff watching *#Scanners* [12] is primed for a cinema screen experience. The perspectives of Nguyen [36] and Abrahams and Kemp [41] focus primarily on the relationship between audiences and the artist or their work, whereas our interests also focus on the relationship, or co-creation between AS and human artists. So, creating artistically sincere TAS for all stakeholders would appear to be a major future challenge for our field.

Lee and See [35] suggest we should not necessarily strive for greater trustworthiness in AS, but rather focus on establishing an appropriate level of trust, one that "describes a relationship that depends on the characteristics of the trustee, the trustor, and the goal-related context of the interaction". Appropriateness of trust is established through matters of calibration between a person's understanding of an AS capabilities; resolution between expectations and capabilities; and the specificity of trust to particular components of an AS and temporal / contextual experience [35].

8 RESPONSIBLE IRRESPONSIBILITY

Finally, we reflect on how the challenges we proposed in this paper align with the Responsible Research and Innovation (RRI) approach that underpins the TAS programme. Within this RRI approach, responsibility is positioned as a means of taking care of the future through actions in the present [42]. How can embracing ambiguous interpretation over clear explanation, failure and improvisation over dependability, and playfulness over caution be deemed to be responsible? We offer three broad responses.

First, we propose that delivering public artworks is a powerful way to engage the public first-hand with the potential of, and challenges arising from emerging technologies such as AS and AI. While other methods such as speculative design fiction [43] can also engage people in reasoning about future technologies, artworks such as those described above give the public personal, up-close, direct, memorable and provocative engagements with new technologies. Experiencing a brain-controlled film or riding a breath-controlled rodeo bull are powerful ways of stimulating a deep personal engagement, as well as drawing in critics and other cultural commentators. In terms of Jirokta et al's [44] "AREA Plus" framework for responsible innovation, the artist may be seen to anticipate a possible future, and, engaging with a public audience, encouraging them (and the researchers who they work with) to do the same, and to reflect on their own positions and commitments. The future implied by the artist's work may be dark or challenging, or hopeful and encouraging, or simply ambiguous. Yet, while intellectually provocative and even sometimes even offensive or otherwise controversial, art is rarely dangerous, certainly not in the way that publicly deploying unproven autonomous vehicles or surgical robots might be. In short, art can provide a powerful and yet safe experimental space within which to responsibly explore issues such as trust in AS and the different potential impacts of automated systems.

Second, art, as part of the wider creative and cultural sectors, is economically, societally and personally important. As a major sector of the economy, artistic applications are worthy of consideration by researchers, but this involves addressing them on their own terms by recognising and respecting their driving goals such as creativity, interpretation and improvisation. Innovation, on the other hand, is generally preoccupied with commercial possibilities or measurable progress in addressing societal challenges. Consider, for example, the widely used (European) assertion that "Responsible Research and Innovation (RRI) is the on-going process of aligning research and innovation to societal values, needs and expectations" [45]. Societally, art provides a mechanism for confronting and challenging the normative thinking of the day and breaking taboos, which often makes it controversial, but which can ultimately be seen as the responsible thing to have done. At a personal level, engaging with art meets basic human values of stimulation, hedonism and self-direction [46]. Sensation-seeking, risk-taking and boundary crossing, even if uncomfortable, may be important for entertainment, personal enlightenment and social bonding [47]. And as Nguyen [36] argues, part of the value of art is in its very diversity and particularity. So, the essential nature of responsibility in relation to art may be more complex and nuanced than it is in relation to technological innovation.

Third, however, we are in no way arguing for a lack of responsibility in approaching artist-led explorations of TAS. Artists and the researchers who collaborate with them need to carefully judge the crossing of boundaries and how they negotiate consent and anonymity and build and maintain trust with their audiences. Artists do not operate in a vacuum, but rather operate within an artworld of funders, venues, curators and critics that hold them accountable for their actions while managing ongoing concerns of safety, liability, accessibility and other important matters (for a richer account of how artists negotiate issues of responsibility and ethics see [48]). One way to articulate our argument is in terms of a call for 'responsible irresponsibility', the idea is that enabling people to artistically play with emerging technologies requires giving them license to exercise a degree of apparent irresponsibility in pushing the boundaries of the technologies and themselves while actually ensuring that they behave responsibly and ethically within the norms and processes of a recognised artworld. The RRI approach itself provides an opportunity for this. RRI's emphasis on engaging with, and being responsive to, stakeholders across society creates a space for dialogue to establish how boundaries can be pushed effectively and responsibly in the particular context of each artist-led exploration. When artists and researchers collaborate, our worlds – including what it means to be responsible – can collide or overlap uncomfortably. This gives rise to a different kind of boundary, which can be very hard to delineate precisely. Ultimately, the playful and irresponsible elements of an experience need to be carefully demarcated within the so-called 'frame' (for artworks) [48] or 'magic circle' [49] (for games) so that we can manage consent and safety. In short, we need to be responsible in enabling people to adopt a bounded form of irresponsibility towards technologies to allow for a richer human experience, even when it appears to fly in the face of conventional wisdom.

9 CONCLUSION

Art is important. It can also be challenging, both at an individual level where it can be difficult to understand or may provoke strong emotions, but also at a societal level where it may confront the received wisdom of the day. Reflecting on a portfolio of artistic experiences involving different kinds of autonomous systems has led us to question conventional wisdom concerning the human experience of autonomous systems: that they should always be explainable, dependable, controllable, and cautious tools. We have explored alternative stances in which AS are experienced as ambiguously open to multiple interpretations, deliberately invoke errors to prompt improvisation,

demand that humans surrender control, are viewed as co-creators, and are playful. In turn, these challenges shed new light on the key matters of trustworthiness and responsibility, leading us to highlight the importance of trust in artistic sincerity and the idea of a playful, if measured, degree of responsible irresponsibility. We argue that these are important considerations or future artistic applications of TAS, but that they might also have wider applicability to domains beyond art if we wish to pursue TAS that engage with humans in a more roundedly human way.

Focusing on the creative and cultural sector, we reflect on the potential impact of a more creative TAS on employment. There is already widespread concern that apparently creative AI may lead to job losses or dislocations with the sector, both for the artists whose works and styles may be emulated or copied by AI, but also for the many animators, illustrators, photographers, musicians and other craftspeople who support them and whose jobs might be replaced. How can we reconcile our call for more creative TAS with such a prospect? The emerging applications of AI in this sector (and machine learning in particular) are fundamentally and essentially creating derivative works. This is not the kind of "creativity" that we are exploring here. The systems frequently embody the tool-oriented values of reliability and human control, and, at least in their more responsible forms, explainability and caution (at least in terms of intellectual property). We are not advocating for TAS to displace more people in creative roles. Indeed, one answer may lie in trying to design AS that stimulate humans to find new forms of creative expression. Can the technology enable the humans involved to be more improvisational, interpretative and playful themselves? Our call to action is perhaps best considered as being to design a more creative human experience of TAS.

We also reflect on the implications of our proposals in other, less overtly creative, sectors, for example in safety critical applications. In this context we reflect briefly on the five challenges suggested here. First, it is clearly provocative – and dangerous – to suggest that a safety critical system might deliberately fail (although safety critical systems are often designed to fail safely rather than arbitrarily). But it is also the case that every autonomous system is itself part of a larger socio-technical system, and safety and tolerance of faults may be achieved at the level of the larger system, even where the smaller system fails. The safety and reliability characteristics of this larger system will necessarily impose constraints on the range of appropriate responses to failure within the system, for example whether improvisation may be acceptable in place of established contingency measures. Second, and somewhat in tension with first challenge, the notion of surrendering to an autonomous system presumes that it is safe to give up a significant degree of agency to the autonomous system – or at least the larger system in which it is experienced. Also, while control may be a reasonable principle for simple goal-directed tasks, loss of control may be more relevant for more nuanced purposes such as revealing challenges and uncertainties. Third, the movement from explainability to interpretation reflects contrasting epistemic commitments. A forensic investigation or a dispute about technical liability may rely on a positivist notion of explainability. On the other hand, more subjectivist pursuits of personal and group meaning are natural territories for interpretation. Fourth, our critique of caution is not an advocacy for harm, but rather a challenging of caution as the only or dominant logic for all design choices. Within the range of the ordinary and everyday risks of being alive (e.g. walking down the pavement), and in the context of a broader system-wide commitment to care, playfulness can have value both in itself and as an axis for innovation and exploration. Finally, in a similar way, the reconceptualisation of autonomous systems as co-creators rather than tools, while being a symbolic renouncing of human superiority, responsibility and control, should be set within a larger context that still recognises the distinct nature of the autonomous system, not least, at the current time, their essential lack of moral agency and "human" rights.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council [grant numbers EP/L019981/1, EP/L015463/1, EP/M02315X/1, EP/G037574/1, EP/G065802/1]. We would like to acknowledge the financial support we received from the Kone Foundation, BFI Film Hub Propeller Scheme, Arts Council England Grants for the Arts, and the EPSRC Telling Tales of Engagement competition. The Digital Score project (DigiScore) is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. ERC-2020-COG – 101002086).

REFERENCES

<bib id="bib1"><number>[1]</number>Laprie, J.C., 1985. Dependable computing and fault-tolerance. Digest of Papers FTCS-15, 10(2), p.124.</bib>
<bib id="bib2"><number>[2]</number>Popat, S. and Whatley, S. eds., 2020. Error, Ambiguity, and Creativity: A Multidisciplinary Reader. Springer Nature.</bib>
<bib id="bib3"><number>[3]</number>Hazzard, A., Greenhalgh, C., Kallionpaa, M., Benford, S., Veinberg, A., Kanga, Z. and McPherson, A., 2019, May. Failing with style: Designing for aesthetic failure in interactive performance. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14).</bib>
<bib id="bib4"><number>[4]</number>Porcheron, M., Fischer, J.E., Reeves, S. and Sharples, S., 2018, April. Voice interfaces in everyday life. In proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).</bib>
<bib id="bib5"><number>[5]</number>Suchman, L.A., 1987. Plans and situated actions: The problem of human-machine communication. Cambridge University Press.</bib>
<bib id="bib6"><number>[6]</number>Shneiderman, B. (1993). 1.1 direct manipulation: A step beyond programming languages. In B. Shneiderman (Ed.), Sparks of innovation in human-computer interaction (p. 17–38). Ablex Publishers.</bib>
<bib id="bib7"><number>[7]</number>Shneiderman, B., & Plaisant, C. (2010). Designing the user interface: Strategies for effective human-computer interaction. Pearson Education India.</bib>
<bib id="bib8"><number>[8]</number>Eigenfeldt, A., Bown, O. and Carey, B., 2015, June. Collaborative Composition with Creative Systems: Reflections on the First Musebot Ensemble. In ICCC (pp. 134-141).</bib>
<bib id="bib9"><number>[9]</number>Vear, C., 2021. Creative AI and musicking robots. Frontiers in Robotics and AI, 8, p.631752.</bib>
<bib id="bib10"><number>[10]</number>Brooks, R. (1987). Intelligence without Representation. Artif. intelligence 47 (1), 139–159.</bib>
<bib id="bib11"><number>[11]</number>Marshall, J., Rowland, D., Rennick Egglestone, S., Benford, S., Walker, B. and McAuley, D., 2011, May. Breath control of amusement rides. In Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems (pp. 73-82).</bib>
<bib id="bib12"><number>[12]</number>Pike, M., Ramchurn, R., Benford, S. and Wilson, M.L., 2016, May. #scanners: Exploring the control of adaptive films using brain-computer interaction. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5385-5396).</bib>
<bib id="bib13"><number>[13]</number>Ramchurn, R., Martindale, S., Wilson, M.L. and Benford, S., 2019, May. From Director's Cut to User's Cut: to Watch a Brain-Controlled Film is to Edit it. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14).</bib>
<bib id="bib14"><number>[14]</number> Benford, S., Ramchurn, R., Marshall, J., Wilson, M.L., Pike, M., Martindale, S., Hazzard, A., Greenhalgh, C., Kallionpää, M., Tennent, P. and Walker, B., 2021. Contesting control: journeys through surrender, self-awareness and looseness of control in embodied interaction. Human–Computer Interaction, 36(5-6), pp.361-389.</bib>
<bib id="bib15"><number>[15]</number>Csikszentmihalyi, M. (1991). Flow, the psychology of optimal experience, steps towards enhancing the quality of life. Harper & Row</bib>
<bib id="bib16"><number>[16]</number>Holzinger, A., "From Machine Learning to Explainable AI," 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Košice, Slovakia, 2018, pp. 55-66, doi: 10.1109/DISA.2018.8490530.</bib>
<bib id="bib17"><number>[17]</number>Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K. and Müller, K.R. eds., 2019. Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700). Springer Natur</bib>
<bib id="bib18"><number>[18]</number>Gaver, W.W., Beaver, J. and Benford, S., 2003, April. Ambiguity as a resource for design. In Proceedings of the 2003 CHI Conference on Human Factors in Computing Systems (pp. 233-240).</bib>
<bib id="bib19"><number>[19]</number>Sengers, P. and Gaver, B., 2006, June. Staying open to interpretation: engaging multiple meanings in design and evaluation. In Proceedings of the 6th conference on Designing Interactive systems (pp. 99-108).</bib>
<bib id="bib20"><number>[20]</number>Boehner, K., Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2005.Affect: From Information to Interaction. In Between sense and sensibility</bib>
<bib id="bib21"><number>[21]</number>Small, C., 1998. Musicking: The meanings of performing and listening. Wesleyan University Press.</bib>
<bib id="bib22"><number>[22]</number>Benford, S., Greenhalgh, C., Hazzard, A., Chamberlain, A., Kallionpää, M., Weigl, D.M., Page, K.R. and Lin, M., 2018, April. Designing the audience journey through repeated experiences. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).</bib>
<bib id="bib23"><number>[23]</number>Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), pp.1-35.</bib>
<bib id="bib24"><number>[24]</number>Skitka, L.J., Mosier, K. and Burdick, M.D., 2000. Accountability and automation bias. International Journal of Human-Computer Studies, 52(4), pp.701-717.</bib>

<bib id="bib25"><number>[25]</number>Suresh, H. and Guttag, J., 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In Equity and access in algorithms, mechanisms, and optimization (pp. 1-9).</bib>
<bib id="bib26"><number>[26]</number>Schwarting, W., Alonso-Mora, J. and Rus, D., 2018. Planning and decision-making for autonomous vehicles. Annual Review of Control, Robotics, and Autonomous Systems, 1, pp.187-210.</bib>
<bib id="bib27"><number>[27]</number>Hughes, B., 2013. Evolutionary playwork and reflective analytic practice. Routledge.</bib>
<bib id="bib28"><number>[28]</number>Brussoni, M., Olsen, L., Pike, I. and Sleet, D.A. 2012. Risky play and children's safety: balancing priorities for optimal child development. International journal of environmental research and public health 9, 9 (2012), 3134–3148.</bib>
<bib id="bib29"><number>[29]</number>Pellegrini, A.D. and Smith, P.K., 1998. Physical activity play: The nature and function of a neglected aspect of play. Child development 69, 3 (1998), 577–598.</bib>
<bib id="bib30"><number>[30]</number>Benford, S., Greenhalgh, C., Giannachi, G., Walker, B., Marshall, J. and Rodden, T., 2012, May. Uncomfortable interactions. In Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems (pp. 2005-2014).</bib>
<bib id="bib31"><number>[31]</number>Marshall, J., Tennent, P., Li , C., Núñez Pacheco, C., , Garrett, R., Tsaknaki, V., Höök, K., Caleb-Solly, P., Benford, S., Collision Design, to appear in Extended Abstracts (Alt.chi) of Proceedings of the ACM Conference on Computer-Human Interaction (CHI 2023), ACM, 2023</bib>
<bib id="bib32"><number>[32]</number>Sturm, B.L. and Ben-Tal, O., 2017. Taking the models back to music practice: Evaluating generative transcription models built using deep learning. Journal of Creative Music Systems, 2, pp.32-60.</bib>
<bib id="bib33"><number>[33]</number>Benford, S., Hazzard, A., Chamberlain, A., Glover, K., Greenhalgh, C., Xu, L., Hoare, M. and Darzentas, D., 2016, May. Accountable artefacts: the case of the Carolan guitar. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 1163-1175).</bib>
<bib id="bib34"><number>[34]</number>Chung, J.J.Y., He, S. and Adar, E., 2022, June. Artist support networks: Implications for future creativity support tools. In Designing Interactive Systems Conference (pp. 232-246).</bib>
<bib id="bib35"><number>[35]</number>Lee, J. D., & See, K. A., 2004, Trust in Automation: Designing for Appropriate Reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392</bib>
<bib id="bib36"><number>[36]</number>Nguyen, C.T., 2021. Trust and Sincerity in Art. Ergo, 8.</bib>
<bib id="bib37"><number>[37]</number>Savery, Richard, Ryan Rose, and Gil Weinberg. "Establishing human-robot trust through music-driven robotic emotion prosody and gesture." In 2019 28th IEEE international conference on robot and human interactive communication (RO-MAN), pp. 1-7. IEEE, 2019.</bib>
<bib id="bib38"><number>[38]</number>Rousseau, Denise M., Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. "Not so different after all: A cross-discipline view of trust." Academy of management review 23, no. 3 (1998): 393-404.</bib>
<bib id="bib39"><number>[39]</number>Stav, Iyar. (2014). Musical Plagiarism: A True Challenge for the Copyright Law. DePaul Journal of Art, Technology and Intellectual Property Law, 25(1), Fall Article 2.</bib>
<bib id="bib40"><number>[40]</number>Tennent, P., Reeves, S., Benford, S., Walker, B., Marshall, J., Brundell, P., Meese, R. and Harter, P., 2012. The machine in the ghost: augmenting broadcasting with biodata. In CHI'12 Extended Abstracts on Human Factors in Computing Systems (pp. 91-100).</bib>
<bib id="bib41"><number>[41]</number>Abrahams, D. and Kemp, G., 2022. Trust and the appreciation of art. Ratio, 35(2), pp.133-145.</bib>
<bib id="bib42"><number>[42]</number> Stilgoe, J., Owen, R. and Macnaghten, P., 2013. Developing a framework for responsible innovation. Research policy, 42(9), pp.1568-1580.</bib>
<bib id="bib43"><number>[43]</number>Sterling, B., Wild, L. and Lunenfeld, P., 2005. Shaping things (p. 133). Cambridge, MA: MIT press.</bib>
<bib id="bib44"><number>[44]</number>Jirotka, M., Grimpe, B., Stahl, B., Eden, G., & Hartswood, M., 2017, Responsible Research and Innovation in the Digital Age. Communications of the ACM. 60. 62-68. DOI:10.1145/3064940.</bib>
<bib id="bib45"><number>[45]</number>Gerber, A., Forsberg, EM., Shelley-Egan, C., Arias, R., Daimer, S., Dalton, G., Cristóbal, A.B., Dreyer, M., Griessler, E., Lindner, R., Revuelta, G., Riccio, A., & Steinhaus, N., 2020, Joint declaration on mainstreaming RRI across Horizon Europe, Journal of Responsible Innovation, 7:3, 708-711, DOI: 10.1080/23299460.2020.1764837</bib>
<bib id="bib46"><number>[46]</number>Schwartz, Shalom H. (1992), "Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries", Advances in Experimental Social Psychology Volume 25, Advances in Experimental Social Psychology, vol. 25, Elsevier, pp. 1–65, </bib>
<bib id="bib47"><number>[47]</number>Lyng, S., 1990. Edgework: A social psychological analysis of voluntary risk taking. American journal of sociology, 95(4), pp.851-886.</bib>
<bib id="bib48"><number>[48]</number>Benford, S., Greenhalgh, C., Anderson, B., Jacobs, R., Golembewski, M., Jirotka, M., Carsten Stahl, B., Timmermans, J., Giannachi, G., Adams, M., Row Farr, J., Tandavanitj, N., & Jennings, K., 2015, The Ethical Implications of HCI's Turn to the Cultural. ACM Trans. Comput.-Hum. Interact. 22, 5, Article 24 (October 2015), 37 pages. https://doi.org/10.1145/2775107</bib>
<bib id="bib49"><number>[49]</number>Tekinbas, K.S. and Zimmerman, E., 2003. Rules of play: Game design fundamentals. MIT press.</bib>