SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

When Being Wrong Is Expensive: Non-Punitive Choice Architecture in Interactive Instructional Media

[1]Nishant Bhuvanesh Trivedi, [2]Dr. Shruti Tiwari

[1]Vadodara, Parul Institute of Design, Parul University. E-mail: nishant.trivedi89029@paruluniversity.ac.in
[2]Vadodara, Parul Institute of Design, Parul University. E-mail: shruti.tiwari25189@paruluniversity.ac.in

Abstract
Interactive instructional media teach through consequence. A learner chooses, the system responds, and the difference between responses is where the learning is supposed to happen. This works because being wrong is informative. It assumes, however, that being wrong is affordable. For some learner groups it is not: where a history of correction, social failure or elevated anxiety makes the experience of error costly in itself, an artefact that dramatises the learner's mistake may teach avoidance of the artefact rather than the skill. This paper develops a conceptual framework for that situation. We first set out the case for choice-based instruction and the case for error, including the productive failure literature and its boundary conditions — notably the finding that younger learners and those with limited prior knowledge do not reliably benefit from failing first. We then characterise failure sensitivity as a population-level design constraint rather than an individual fragility, drawing on the elevated prevalence of anxiety in autistic children and on the errorless teaching tradition in developmental disability education. From this we propose a four-part typology of branch topologies — punitive, corrective, differentiated and equivalued — and identify the fidelity–safety trade-off as the central design tension: consequence carries information, and removing it removes information along with the harm. We specify five conditions under which non-punitive branching should be preferred, four design moves that preserve informativeness without punishment, and the objections that a designer adopting this position must be able to answer. The framework is illustrated briefly with reference to one instantiation, an animated social skills curriculum for autistic children, but is stated to apply wherever choice-based instruction meets a failure-sensitive learner group. This is a conceptual contribution; no empirical claims are made and none should be inferred.

Keywords: instructional design; interactive narrative; branching; learner agency; failure sensitivity; autism spectrum condition; educational technology

1. Introduction

Every branching instructional artefact makes a decision that its designers rarely state out loud: what happens when the learner picks the option the designer would not have picked.

The default answer, inherited from games and from most of what we know about learning, is that something bad happens. The character fails. The score drops. A message says try again. The badness is not incidental — it is the pedagogy. Consequence discriminates between options, and discrimination is what makes the choice a choice rather than a button press. Remove it and the interaction appears to become decorative.

This paper is about the cases where that default is wrong, and about what to do instead without simply making the interaction decorative.

The argument turns on a distinction between two things that are usually conflated. Being wrong carries information: the learner discovers that a particular action does not achieve a particular goal. Being made to feel wrong carries something else: an affective experience of exposure, correction and failure. In most learners and most contexts the second is a tolerable side effect of the first, and can even be useful. But the two are separable in principle, and there are learner groups for whom the second is expensive enough that it dominates the first. We call this property **failure sensitivity**, and we treat it as a characteristic of populations, not of individuals. This matters, because the individual framing invites a remedial reading — this child is fragile, and needs protecting — whereas the population framing states a design requirement: if the artefact will be used by a group in which a large proportion carries elevated cost for the experience of error, then the artefact's branch structure is a design variable that has to be reasoned about rather than inherited.

Three questions organise the paper:

Under what conditions does the informational value of instructional error fail to justify its affective cost?

What branch topologies are available to a designer, and what does each one buy?

If negative consequence is withheld, what carries the discriminative information that made the choice meaningful?

The third question is the hard one, and most of the paper is about it. A framework that answers only the first two amounts to an argument for removing consequence, which would be an argument for a worse artefact.

We illustrate with reference to UD (Understanding Duo), an animated social skills curriculum for autistic children whose design is reported separately [AUTHOR: cite the companion design paper here once it has a DOI or preprint identifier]. The illustration is short and deliberately subordinate. The framework is the contribution; UD is one instantiation of it, and the reader should be able to reject the instantiation while retaining the framework.

2. The case for choice

The argument for building choice into instructional media has three independent strands, and it is worth separating them because they imply different things about what a good choice looks like.

Motivational. Self-determination theory holds that autonomy, competence and relatedness are basic psychological needs, and that conditions supporting their satisfaction promote self-motivated, high-quality engagement while conditions thwarting them undermine it (Ryan & Deci, 2000). A choice the learner experiences as genuinely theirs is autonomy-supportive. A choice that is nominally open but whose correct answer is signalled by tone, framing or interface emphasis is not — it is a compliance test wearing the costume of a choice, and learners detect this readily.

Cognitive. Instruction that supplies the answer never exercises the step in which the learner appraises a situation and selects among candidate actions. In domains where competence consists precisely in that selection — social behaviour is the clearest case, but clinical reasoning, negotiation and ethical judgement are others — an artefact that always specifies the correct action trains retrieval rather than judgement.

Structural. Petousi et al. (2022), working on digital storytelling for history education, treat the decision point as the site where the pedagogically interesting activity happens rather than as a mere navigation device — the branches exist to make the moment of deciding matter. This is a claim about where to concentrate design effort, and it implies that decision points should be few, deliberate and consequential rather than scattered.

These three arguments agree that choice is valuable. They disagree, subtly, about why, and the disagreement resurfaces in Section 6.

3. The case for error, and its boundaries

The strongest argument against the position developed here comes from the productive failure literature, and it deserves to be stated at full strength rather than caricatured.

Productive failure sequences problem solving before instruction: learners attempt a problem they are likely to fail, generate multiple partial and incorrect solutions, and only then receive canonical instruction. The claim is that the initial failure activates prior knowledge, surfaces the gaps and inconsistencies in it, and prepares the learner to encode the subsequent instruction more deeply. The characteristic empirical signature is that direct instruction and productive failure perform comparably on recall, while productive failure outperforms on transfer to novel problems. Error, on this account, is not a cost to be minimised but the mechanism itself.

If that is right in general, an artefact that withholds error is withholding the active ingredient.

Two boundary conditions constrain the generalisation, and both are load-bearing for the present argument.

The first is prior knowledge. The mechanism runs through activation and differentiation of what the learner already knows; a learner with little relevant prior knowledge generates few solutions and differentiates little, and the failure is simply failure.The second is age, and here the meta-analytic evidence is pointed. Sinha and Kapur (2021) synthesised 166 experimental comparisons across 53 studies of problem-solving-before-instruction against instruction-before-problem-solving. Overall they found a moderate effect favouring the failure-first sequence (Hedges' $g = 0.36$, 95% CI [0.20, 0.51]), rising to between 0.37 and 0.58 where the design adhered closely to productive failure principles. But grade level moderated the result, and for younger learners — roughly second to fifth graders — the trend reversed, with effect sizes favouring the instruction-first sequence instead. This is not a refutation of productive failure; it is a statement about where its benefits have been demonstrated, and the age band in which they have not is the one that concerns us here. The learners for whom the failure-first case is weakest are precisely the primary-age learners who are the audience for most social and behavioural instructional media.

A third point is definitional rather than empirical. Productive failure concerns failure at a task — a mathematics problem, a physics prediction — where the failure is impersonal and the domain is one in which nobody expects a novice to succeed. Social behaviour is not like this. A child who chooses the wrong thing to say to a classmate has not failed at a problem; they have failed at being a person in a room with another person, which is a category of failure they may have extensive and unhappy prior experience of. The productive failure literature does not license transporting its findings into that category, and we should not assume it does.

4. Failure sensitivity as a design constraint

We now characterise the population property that motivates the framework.

Anxiety. Anxiety disorders are markedly elevated in autistic children and young people. Van Steensel, Bögels and Perrin (2011), meta-analysing 31 studies covering 2,121 young people, found that 39.6% met criteria for at least one anxiety disorder, with specific phobia most common at 29.8%, followed by obsessive–compulsive disorder at 17.4% and social anxiety disorder at 16.6%. Van Steensel and Heeman (2017), in a larger synthesis of 83 articles, confirmed elevated anxiety levels relative to typically developing children and found the difference to increase with IQ — which matters here, because higher-IQ autistic children are disproportionately the users of screen-based instructional media.

Accumulated correction. Beyond diagnosable anxiety, children whose behaviour is a routine target of adult intervention accumulate a history in which their spontaneous social actions are frequently followed by correction. An instructional artefact that reproduces that contingency is not neutral; it is one more instance of a pattern the child already has.

Existing practice already assumes this. Behaviour-analytic instruction in developmental disability has long used errorless teaching procedures — prompting, stimulus fading, graded choice — precisely to minimise the occurrence of errors during acquisition, on the twin grounds that errors can become established through repetition and that error-heavy teaching produces avoidance and emotional responding. Whatever one thinks of that tradition's other commitments, it represents a substantial body of practice built on the premise that error suppression is sometimes instructionally correct. The position developed here is less radical than it may appear; it extends an existing practice principle into a medium that has not previously applied it.

The conjunction of these three gives a population-level constraint. It is not that any particular child cannot tolerate being wrong. It is that in a group where roughly two in five carry a diagnosable anxiety condition and most carry a history of correction, the expected cost of a punitive branch is high enough to enter the design calculus explicitly.

5. A typology of branch topologies

We distinguish four topologies. The typology is offered as a design vocabulary; its value is in making the choice between them visible, since at present most artefacts adopt the first two without deliberation.

Type 1 — Punitive. At least one option leads to an outcome that is bad for the character and legible as bad for the learner: ridicule, exclusion, failure, an unrecoverable state. Discriminative information is maximal. Affective cost is maximal. This is the default in entertainment branching and in much gamified instruction.

Type 2 — Corrective. Non-preferred options are blocked, reversed, or met with try again. This appears gentler than Type 1 but is arguably worse for autonomy: it reveals that the choice was never real, which is the specific thing self-determination theory identifies as motivationally corrosive (Ryan & Deci, 2000). It also destroys the narrative continuity that made the branch interesting.

Type 3 — Differentiated. All options produce complete, constructive, recoverable outcomes, but the outcomes differ in how well they serve the character's stated goal. The learner observes that one action got the character closer to what the character wanted. No option produces humiliation and none is blocked. Discriminative information is preserved; affective cost is low.

Type 4 — Equivalued. All options are valid solutions and differ only in texture. The learning is not which action is better but this goal admits several actions, and you may prefer one. This is not a degenerate case: for goals that genuinely have multiple valid solutions — greeting someone, expressing enthusiasm, declining an invitation — Type 4 is the accurate representation and Type 3 would misrepresent the domain.

A curriculum built entirely from Type 4 would teach that all social actions are equivalent, which is false. One built entirely from Type 3 would make every decision a test. The design question is the distribution across a body of material, and it should be specified per unit and justified.

6. The fidelity–safety trade-off

Here is the objection the framework has to survive.

Real social errors do sometimes carry real cost. A child who pushes to the front of a queue may be told off. A child who interrupts repeatedly may be excluded. If an instructional artefact systematically represents a world in which no social action carries negative consequence, it is teaching something false, and it may be teaching it to learners who are already less well equipped to infer the omitted information from elsewhere.

We do not think this objection can be dissolved. It can only be traded against, and the trade should be made explicitly.

The trade is favourable when the following conditions hold. We state them as conditions rather than as a rule because a designer facing a different combination should reach a different conclusion.

C1 — Elevated failure cost. The intended learner group carries documented, population-level elevated cost for the experience of error, whether through anxiety, correction history, or both.

C2 — Voluntary use. Continued engagement with the artefact depends on the learner's willingness rather than on compulsion. Where use is voluntary, an artefact that produces aversive experiences loses the learner entirely, and an artefact that has lost its learner has an effect size of zero regardless of its pedagogy.

C3 — Acquisition rather than fluency. The instructional purpose is initial acquisition of a repertoire, not refinement of an existing one under realistic pressure. Consequence exposure is more defensible at the fluency stage.

C4 — An external channel for the omitted information. Something other than the artefact — an adult, a subsequent phase, a later unit — is available to convey what the punitive branch would have conveyed. Without this, C1 to C3 are insufficient, and the framework should not be applied.

C5 — Domain admits plural valid actions. Where the domain genuinely has one correct action and the rest are dangerous — the obvious case is physical safety — a non-punitive representation is a misrepresentation and the trade fails.

C4 deserves emphasis because it is the condition most often assumed and least often designed for. Withholding negative consequence creates an information debt. If nothing in the surrounding system pays that debt, the framework is not a design position but an evasion.

7. Four design moves

Given C1–C5, what carries discriminative information in the absence of punishment? Four moves, in rough order of strength.

M1 — Goal-relative differentiation. Give the character an explicit, stated goal at the outset. Discriminate options by distance from that goal rather than by valence. Waiting got him to the counter; leaving the queue meant he did not get the thing he wanted is discriminative without being punitive, because the negative content attaches to the goal state rather than to the learner's worth.

M2 — Counterfactual availability. Let the learner return to the decision point and view the outcome of an option they did not take. This is the single most important move, because it converts the branch from a commitment into a comparison. The learner can hold two outcomes side by side, which is exactly the discrimination a punitive branch was supposed to produce, obtained without anyone having to fail. The framing is load-bearing: see what happens if preserves exploration, try again collapses the design back to Type 2.

M3 — Perspective disclosure. Show the same moment from another character's viewpoint, including what that character felt. This carries social-consequence information — the other person was confused, or was waiting — without staging the learner's action as an offence.

M4 — Non-evaluative reflection. Close the unit with a reflection that names what happened without ranking the learner. The distinction is between waiting worked well there and you chose well. The first is about the world; the second is about the learner, and re-establishes the evaluative frame the topology was chosen to avoid.

M2 does most of the work. A designer forced to implement only one of these should implement M2.

8. Illustration

UD (Understanding Duo) is an animated social skills curriculum for autistic children in which each episode follows a fixed ten-beat structure, three beats of which form the interactive core: an older near-peer guide character asks the viewer what the younger protagonist should do, three options are presented as image, audio and text simultaneously, the chosen action plays out to a non-punitive consequence, and the child may then view an unchosen branch side by side with the one they took. Its design is reported in full separately.

Read through the present framework, UD instantiates the position as follows. C1 holds by population. C2 holds, since use is voluntary and a child who dislikes an episode simply stops. C3 holds, since the curriculum targets acquisition. C4 is addressed through an adult support layer accompanying every episode, which gives parents and teachers the naturally occurring opportunities to prompt the skill and the language to use — this is UD's attempt to pay the information debt, and it is the component most in need of empirical testing. C5 is handled by distribution: the curriculum's safety-domain material is specified differently from its foundational-social material, because the C5 condition fails there. On topology, UD uses Types 3 and 4 and prohibits Types 1 and 2. Its opening episode is Type 4 — three ways of greeting a peer, all of which work — chosen deliberately because a first episode that discriminated between options would establish the evaluative frame the curriculum is trying not to establish. Later episodes are Type 3.All four design moves are present. M2 is implemented as a dedicated beat in the episode structure rather than as an optional feature bolted to the interface — the comparison is part of what an episode is, which is the difference between a design commitment and an affordance. Its framing is see what happens if, not try again.

We offer this as an existence proof that the framework is buildable at curriculum scale, not as evidence that it works. No efficacy claim attaches to it.

9. Objections

"This is coddling, and children need to learn that actions have consequences." The framework does not remove consequence; M1 preserves it in goal-relative form. What it removes is the staging of the learner as having failed. These are different, and conflating them is the conceptual error the paper exists to correct. A child who sees that leaving the queue meant not getting the thing has learned about consequence.

"Children taught in a consequence-free environment will be unprepared for a world that has consequences." This is C4 restated, and it is the objection we take most seriously. It is an argument for building the external channel, not against the topology. It is also an empirical question that the framework generates rather than answers, and we flag it in Section 10 as the first thing worth testing.

"You are asserting failure sensitivity rather than demonstrating that it changes learning outcomes." Correct. The anxiety prevalence data establish that the population differs; they do not establish that the difference changes what a punitive branch does. The framework rests on a plausible inference, and we would rather state that plainly than dress the inference as a finding.

"Type 4 branches make the interaction pointless — any answer works, so why choose?" Because the learning content of a Type 4 episode is not which but that: that the goal admits several actions and the learner may have a preference. For a learner who has been taught that social situations have single correct answers, this is substantive rather than trivial content. It is also the only topology under which the learner's own preference is treated as legitimate information rather than as noise.

"This is a repackaging of errorless learning." It shares errorless learning's premise and differs in what it does with choice. Errorless teaching minimises error by constraining the learner's opportunity to respond incorrectly, typically through prompting and fading. The present framework leaves the response space fully open — three genuinely available options.

10. A research agenda

The framework is a set of conjectures. The following would test them, in rough order of tractability.

Does topology affect continued engagement? Compare Type 1 and Type 3 versions of otherwise identical material in a failure-sensitive group, with voluntary continuation as the outcome. This is the framework's most basic claim and the easiest to test.

Does M2 recover the discriminative information? Compare learners who use counterfactual replay against those who do not, on a measure of whether they can articulate why one option served the goal better.

Does C4 matter as much as we claim? Compare the same non-punitive artefact with and without an adult extension layer, on transfer to untrained contexts. If the difference is negligible, C4 is overstated and the framework should be revised.

Where is the boundary of failure sensitivity? The framework is derived from autism but stated generally. Whether it applies to learners with anxiety presentations, with histories of educational failure, or in adult retraining contexts is unknown and worth establishing.

Does a non-punitive curriculum leave learners less prepared for real consequence? The central objection, and the hardest study to run well.

None of these is answered here. Stating them is part of the contribution: a conceptual framework that generates no testable disagreement is not doing useful work.

11. Conclusion

Interactive instructional media inherited their branch structure from entertainment, where being wrong is the point. That inheritance goes unexamined in most educational applications, and for most learners it is harmless. For learners whose experience of being wrong is already expensive, it is not.

We have argued that the informational content of error and the affective experience of error are separable; that four branch topologies are available and most designers use two of them by default; that the trade between ecological fidelity and psychological safety is real and should be made under stated conditions rather than by habit; and that counterfactual availability is the design move that lets a designer withhold punishment without withholding information.

The framework is conceptual. It has not been tested, and the paper should not be read as claiming that non-punitive branching produces better learning than the alternative. What it claims is narrower and, we hope, more useful: that this is a design decision, that it has been made unreflectively, and that there is a principled basis on which to make it differently.

References

1. Nejati, V., Peyvandi, A., Nazari, N., & Abadi, F. (2024). The effectiveness of social training in individuals with autism spectrum disorder (ASD): A systematic review and transfer analysis. Scientific Reports, 14, Article 32131. https://doi.org/10.1038/s41598-024-83953-9
2. Petousi, D., Katifori, A., Servi, K., Roussou, M., & Ioannidis, Y. (2022). History education done different: A collaborative interactive digital storytelling approach for remote learners. Frontiers in Education, 7, Article 942834. https://doi.org/10.3389/feduc.2022.942834
3. Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68–78.
4. Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. Review of Educational Research, 91(5), 761–798. https://doi.org/10.3102/00346543211019105 [AUTHOR: confirm volume, issue and page range at the publisher record before submission.]
5. van Steensel, F. J. A., Bögels, S. M., & Perrin, S. (2011). Anxiety disorders in children and adolescents with autistic spectrum disorders: A meta-analysis. Clinical Child and Family Psychology Review, 14(3), 302–317. https://doi.org/10.1007/s10567-011-0097-0
6. van Steensel, F. J. A., & Heeman, E. J. (2017). Anxiety levels in children with autism spectrum disorder: A meta-analysis. Journal of Child and Family Studies, 26(7), 1753–1767. https://doi.org/10.1007/s10826-017-0687-7 [AUTHOR: confirm issue and page range at the publisher record before submission.]

Appendix A — Branch topologies at a glance

Topology	Non-preferred option leads to	Discriminative information	Affective cost	Autonomy
1 Punitive	Ridicule, exclusion, unrecoverable state	High	High	Preserved
2 Corrective	Block, reversal, try again	Moderate	Moderate	Destroyed
3 Differentiated	Complete, constructive, further from goal	Moderate–high	Low	Preserved
4 Equivalued	Complete, constructive, equally valid	Low (by design)	Minimal	Preserved

Appendix B — Conditions and moves

	Statement	Fails when
C1	Population carries elevated failure cost	Learner group is unselected
C2	Use is voluntary	Participation is compelled
C3	Purpose is acquisition	Purpose is fluency under pressure
C4	External channel pays the information debt	Artefact stands alone
C5	Domain admits plural valid actions	One action is correct, others dangerous
M1	Goal-relative differentiation	No stated character goal
M2	Counterfactual availability	Replay framed as correction
M3	Perspective disclosure	—
M4	Non-evaluative reflection	Reflection addresses the learner