Prototyping Kant-inspired Reflexive Game Mechanics

Mark J. Nelson
Center for Computer Games Research
ITU Copenhagen
Copenhagen, Denmark
mjas@itu.dk

ABSTRACT

Immanuel Kant's categorical imperative, stating (in one formulation) that one ought to always act according to a maxim that can be made universal law, is tempting to proceduralize, in the form of a game that literally turns actions into universal laws. This paper explores difficulties that initially arise in translating that idea to a game design: some of which been covered in the philosophy literature, and others of which relate to the difficulties in defining what constitutes a proper rule induction. Then, it discusses several much less lofty, but practical, prototypes that explore what I take to be the formal game mechanics underpinning the idea: reflexive game mechanics where breaking a rule implies the free breaking of that rule for the rest of the game. By analyzing these prototypes, I attempt to determine if these prototypes result in either an interesting game mechanic (taken on its own) on the one hand, or a compelling representation of Kantian morality on the other hand, reaching mixed conclusions.

1. INTRODUCTION

A philosophical thought experiment describes a hypothetical world or situation in order to gain insight into a philosophical question. Games can represent and provide a virtual experience of imagined worlds and situations. Thus we might want to ask of any given philosophical thought experiment: can this be instantiated as a playable game? Furthermore, is anything gained by doing so, either for philosophy or for game design?

In this paper I follow that approach by attempting to derive game mechanics from Immanuel Kant's categorical imperative. After I find a number of difficulties with directly representing the thought experiment implied by the categorical imperative, I shift towards what I claim is the core idea from a gameplay perspective, reflexive game mechanics where rule-breaking changes the rules of the game. This idea is directly implementable, but somewhat removed from the original philosophical idea. I discuss two small game prototypes built to use those mechanics, intended to understand how the mechanics operate. Finally, I relate these prototypes back to the original goal of procedurally representing the categorical imperative, and compare to an alternative approach recently investigated by Togelius [6].

2. THE CATEGORICAL IMPERATIVE

The categorical imperative is the core of Kant's system of deontological ethics, claiming to give a criterion by which we can judge actions as inherently ethical or unethical. He gives several formulations, but of interest here is the universal-law formulation, which comes closest to directly setting up a thought experiment: "act only according to that maxim whereby you can at the same time will that it should become a universal law".

One interpretation of this formulation, due to Christine Korsgaard [3], is that an action is unethical if you cannot consistently imagine it as a universal law without defeating the purpose of taking the action in the first place:[1]

> The contradiction is that your maxim would be self-defeating if universalized: your action would become ineffectual for the achievement of your purpose if everyone (tried to) use it for that purpose. Since you propose to use that action for that purpose at the same time as you propose to universalize the maxim, you in effect will the thwarting of your own purpose.

The standard example is lying: If everyone lied when it benefited them, lying would cease to be effective, since its effectiveness depends on the fact that people believe you, and nobody would believe you if it were standard practice to lie whenever it benefited you.

3. THE CATEGORICAL IMPERATIVE AS GAME MECHANIC

In a game version of the thought experiment, universal laws are game rules. In particular, they're the rules that govern how non-player characters (NPCs) behave. So if a player lies, from that point onwards lying will be something other characters can do, and similarly, believing lies will be something that they won't do anymore—literally realizing the

[1]This is what Korsgaard labels the "practical contradiction" interpretation.

possible world that the thought experiment implies. This approach avoids having to actually algorithmically determine what constitutes a "contradiction": it realizes the practical contradiction *in practice* by making it the new rule of the simulation, and leaves it to the player to determine which rule changes ended up counting as practical contradictions.

In other words, rather than spitting out "that was a moral action" vs. "that was an immoral action" judgments about the player's actions, this approach would implement the universalization itself as a mechanic. Then, if the player takes actions that would result in a practical contradiction—i.e. universalization would thwart the action's purpose—then that gets represented in the game world by universalization in fact happening to thwart the action's purpose. We don't even necessarily need to actually *know* what the player's purpose in taking a particular action was with such an approach.

An induction difficulty remains. The player is not actually giving us maxims, but taking actions, from which we need to abstract which maxims we assume them to be acting by—maxims that could be at any level of granularity. If the player lies about the price of bread, does this mean that lies about bread are now universal? That all lies are universal? Or something odd, like: all bread now has the price the player said it did? Obviously some abstractions get to the relevant point more than others, but it's not clear how we can automatically infer them. Some simplified method would need to be used, which might be tailored depending on the character of the game. A flippant, highly caricatured game could do well with an abstraction method that extracts grossly overbroad maxims from everything the player does; a more sober game might want a more conservative method. Fortunately, we don't have to solve this for every possible action, since games can also constrain the types of actions players have available.

A second, particularly difficult problem to solve is how to interpret the universalization itself. What, in fact, would happen if this maxim were universalized? Some parts are fairly easy, such as simply letting the other characters do everything that the player seems to have implied was acceptable to do. However others, such as inferring that the other characters shouldn't believe lies once lying becomes a maxim, require quite a bit of common-sense reasoning—a highly nontrivial AI problem. Thus a workable game would require quite careful design to set it in a sufficiently abstracted world to be tractable (not AI-complete) to implement.

One possibility would be to go for a small world with a small number of actions, where the game author can fully hardcode all the relationships, implications, and sensible abstractions manually, avoiding the need for any of this induction. In a sense that's the approach I took, but took the simplification one step further: rather than investigate the categorical imperative at all as a first step, investigate this formal mechanic of player-doing-something-implies-a-new-rule, in a simpler setting where it's clearly how to implement it. The hope is that if we learned something about that kind of reflexive mechanic in general, it would allow us to better understand how to approach a similar mechanic in the context of moral systems (or, alternately, just end up with an interesting mechanic).

More specifically, I investigate starting with a fixed set of rules, and then adding a mechanic that says: when you break a rule, a new rule is induced that makes your action legal (also, how the induction proceeds is simple and hardcoded). Then gameplay continues with the new rule as part of the game.

4. RULE-BREAKING PROTOTYPES

In the first experiment, take *Pac-Man*, and add one rule-breaking possibility: when you hold down a button, you can bulldoze your way through walls. Once you do so, the wall stays gone for the rest of the game, and the hole can be used by the enemy ghosts as well. What effect does this have? Knocking your way through walls can be useful in a pinch, to escape dead-ends. But doing it too much turns the maze into swiss cheese and makes it nearly impossible to actually avoid the ghosts. This actually appears to get at a little of our original goal, showing the possible negative consequences of rule-breaking. There's also a nice visual-representation aspect, where breaking the rules too much literally turns the maze into damaged-looking ruins of the original level. The wall-breaking *Pac-Man* game gets only indirectly at game rules, though; it specifically modifies the level configuration.

Consider a second experiment, this time with a bit more rule-induction. Start with chess, but add the possibility to move pieces in a way that breaks the usual rules of chess. If you move a pawn in a way that would be illegal for a pawn, but would be legal for a bishop, then henceforth all pawns, of both players, can move like bishops (in addition to still being able to move like pawns). More specifically, if a piece attempts to move somewhere that isn't legal for its type, we check if it would be legal for any other type, in the following order of preference: pawn, king, knight, bishop, rook (queen is never needed, since any move that a queen could make could've also been made by either a bishop or a rook). Then we add the induced type's movement abilities to the abilities of the offending type. Figure 1 shows an example.[2]

What effects does this have? Primarily, it makes the games really short and very difficult to play. Contemplating the effects of a move in chess typically requires you to mentally project possibilities a few moves into the future. But now every move includes the possibility of a piece gaining new types of movement, so within a few moves into the future, pretty much anything is possible. It would be interesting to determine if there are dominant strategies in such a game. And perhaps, with an admittedly large stretch, this outcome conveys the ethical judgment that rampant rule-breaking quickly degenerates into hedonistic chaos, where everyone can do anything, and order and structure disappears.

Note that this chaos happens despite the fact that rule-breaking is in a sense already quite strongly restricted. Pieces cannot move in *any* possible way, but only in a way that would be legal for at least one other kind of piece; for exam-

[2]A Java-applet version of this game can be played at `http://www.kmjn.org/notes/reflexive_rules.html`.

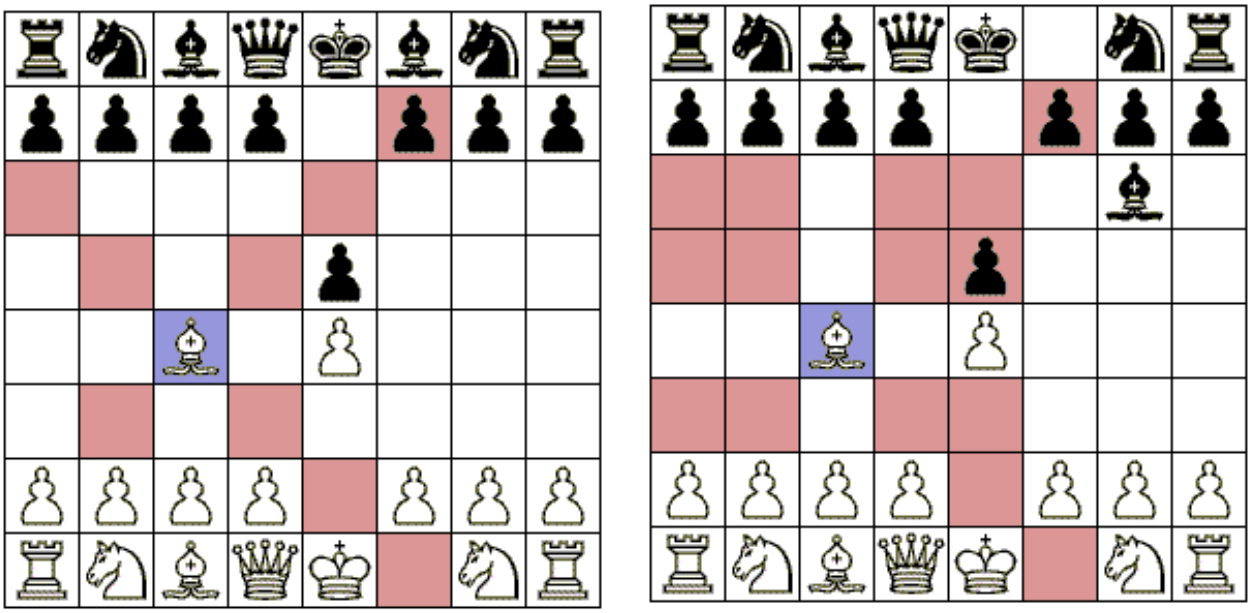

Figure 1: A chess game with reflexive mechanics. On the left, the white bishop is shown being able to move in a normal bishop pattern. On the next turn, the black bishop breaks the normal rules and moves like a knight. The game induces that this would be legal if all bishops could move *either* like bishops or like knights. Henceforth, all bishops in the game can indeed move either like bishops or like knights, as illustrated in the right panel showing the white bishop's new options.

ple there's no way to directly teleport to kill the king on the first move. And rules other than movement are not changed either; players cannot resurrect their dead pieces, add new pieces to the board, or make multiple moves in one turn. They can specifically violate only movement rules, and even then only in restricted ways. But it must be admitted that this modified chess isn't a very fun game.

One attempt at a fix: turn rule-breaking into a limited weapon that can be used some number of times a game, sort of analogous to the powerful bombs in some arcade games that you can use once or twice; but not all the time. This is still mentally more complex than regular chess, because projecting possible future moves and counter-moves requires you to consider the possibility that a piece will break the rules and gain new abilities at any time. But at least you can think of that as the opponent deploying their rule-breaking special move, which can only happen a limited number of times. Actually, to make it manageable, the limit in my experiments has mostly been set at one: each player can break movement rules, thereby inducing new abilities onto a type of piece, exactly once per game. I'm not sure how successful this is as a game still, but it's an interesting dynamic I'd like to investigate further.

5. DISCUSSION

The primary thing that stands out about these prototypes is that the major design issue with a rule-breaking mechanic is to handle the sheer havoc it can wreak. When rule-breaking produces a new permissiveness in the rules, the possibility space of gameplay continually increases, and tends towards unstructured chaos. Indeed, without *any* limitations, it's not even clear that the thought experiment is coherent: if a game has some nominal rules, but none of them have to be followed ever, it's not clear in what sense it ever "had" those rules to begin with.

One possibility is to turn the common chaotic result into a game embodying a convincing procedural rhetoric, where havoc being wreaked by rulebreaking is the *point* of the game, a sort of inevitable failure condition, in the vein of some of the "rhetoric of failure" games designed by Gonzalo Frasca or Ian Bogost [1, 2], in this case intended to highlight the chaotic unworkability that must ensue if rampant rulebreaking is allowed to take place. I haven't seriously explored that avenue, but it's certainly a possibility.

The other possibility is to control the added complexity and game-breaking potential, so it becomes more of a "normal" mechanic adding an interesting twist to the game, rather than a game-smashing mechanic that always results in a chaotic mess. The limit I added to the chess prototype, of one rulebreak per game, is one simple attempt to take that route, though not an entirely satisfying one, since adding a numerical limit to rulebreaking has somewhat of a arbitrary band-aid feel to it (on the other hand, those kinds of limits are fairly traditional gameplay devices).

5.1 Relation to Kantian ethics

An unexpected effect of this series of thought experiments and prototyping is that, when related back to the original Kantian thought experiment, we've in a certain sense, at least when filtered through an intelligent player, ended up close to the *opposite* of the spirit of the categorical imper-

ative. In the philosophical version, the idea was to judge actions by their universalizability: a universalizable action is moral, while one that would "break things" when universalized, by causing practical contradictions, isn't. But in the break-the-rules-once formulation, the goal is actually to actively *try* to find the instance of rule-breaking that is *least* universalizable. Your rulebreaking will be universalized after you do it—once you move your bishop like a knight, all bishops will be able to move like knights—so you gain advantage primarily on the first rule-breaking, when you take a forbidden action that was not at the time universal, and gain an advantage by the temporary asymmetry.

The discrepancy here is due to the time delay. Kant envisions an atemporal (or perhaps retroactive) thought experiment: when someone lies (for example) they imagine a world where everyone has always lied, which produces a contradiction. But in a temporally dependent game implementation, which universalizes actions only when taken, regularly lying is not a feature of the world *up until the first time you do so*. When you lie in a world where lying isn't common, it's effective the first time, and only subsequently are further attempts to gain advantage by lying thwarted.

The temporal aspect caused by universalizing at a point in gameplay causes the entire focus to be shifted towards maximizing the usefulness of your one "break the traditional rules of morality" card. That leads to a focus and strategy that's the opposite of trying to act in a way that's univeralizable. That doesn't necessarily mean the mechanic is broken—it might produce interesting gameplay nonetheless—but this seemingly minor shift, an implementation detail to adapt the thought experiment to the linear-time causal necessities of a videogame, turns out to have fairly serious consequences for the attempt at representing the motivating thought experiment.

5.2 Alternative approach

The prototypes here can be seen as universalizing axioms only in a subtractive direction: we start from a *status quo* representing traditional norms, and then induce new permissivity once the player has taken a previously-banned action. That direction is not necessarily required by Kantian ethics, though many of Kant's own examples (like the lying one recounted here) do have that conservative flavor.

An alternative, explored in similarly motivated but independent[3] research by Togelius [6], is to start with *no* rules, and use player actions to induce new rules in a purely constructive manner, attempting to synthesize new universal laws to explain why their actions took place. That is, while the prototypes in this paper attempt to determine which laws the player has repealed by violating them, Togelius attempts to determine which new laws the player has promulgated by following them.

5.3 Future work

As regards future research, an obvious but intriguing approach would be to experiment with various combinations of the two conceptions of how player actions should implicitly revise rules, allowing player actions to both add and remove rules based on some kind of more general induction scheme. An additional question to ask might be whether *all* gameplay rules and player actions need to participate in this mechanic; after all, in normal life, we often think of only a subset of actions as having ethical import.

On the technical side, a closer connection to questions of rule induction in artificial intelligence in general could be investigated; the ambiguity that causes a problem here, of what general principles are implied when we see a specific example of an action, is a familiar one in machine learning, theory revision, and other areas.

From a philosophical perspective, admittedly these experiments have drifted rather significantly from the original motivation, towards abstract rule spaces. Returning to the goal of elucidating the categorical imperative in a thought experiment might be best done by actually removing the induction issue and hand-coding the induction rules in a smallish world, so as to put the focus back on the ethical tradeoffs in some kind of toy world.

[3]We were each quite surprised to find that the other had independently been prototyping games attempting to encode a notion of Kantian ethics.

Acknowledgments

Some of the ideas in this paper were originated in discussions with Michael Mateas and Ian Bogost in 2006–07, spurred by Bogost teaching a class on "a potential debate in contemporary videogame studies that has enjoyed little discussion: that between play and representation".[4] The idea of attempting to use a take on Kantian ethics in a game grew out of a survey of "morality scoreboard" systems used in popular games such as *Black & White*, and considering possible alternatives [4, 5].

More recently, I've benefitted greatly from discussions with Julian Togelius and Pippin Barr on the subject.

6. REFERENCES

[1] I. Bogost. *Persuasive Games: The Expressive Power of Videogames*. MIT Press, 2007.

[2] G. Frasca. *Play the Message: Play, Game and Videogame Rhetoric*. PhD thesis, IT University of Copenhagen, 2007.

[3] C. M. Korsgaard. Kant's formula of universal law. *Pacific Philosophical Quarterly*, 66(1–2):24–47, 1985. Reprinted in *Creating the Kingdom of Ends*, Cambridge University Press, 1996.

[4] M. J. Nelson. Moral calculus in videogames: Some ways it is and might be done. `http://www.kmjn.org/notes/moral_calculus_in_videogames.html`, 2006. Online essay.

[5] M. J. Nelson. The morality systems of Black and White and Fable: A review. `http://www.kmjn.org/notes/moral_systems_bw_fable.html`, 2006. Online essay.

[6] J. Togelius. A procedural critique of deontological reasoning. In *Proceedings of the 5th Conference of the Digital Games Research Association (DiGRA)*, 2011.

[4]`http://www.bogost.com/teaching/game_design_and_analysis.shtml`