Poama, A., dans Origgi, G. (éd.). *Passions Sociales*. Paris : PUF, 2019.

PUNITION

Le terme *punition* est exposé à une double équivoque. La première équivoque relève d'une ambiguïté. En français, le terme peut évoquer soit l'imposition d'une *sanction* sur une personne, soit l'impact psychologique (ou, parfois, physique) négatif que la sanction exerce sur la personne sanctionnée. L'ambiguïté ici découle d'une potentielle confusion entre punition comme sanction dont les effets psychologiques (ou physiques) restent conceptuellement non pertinents et punition comme rétribution ou affliction. La deuxième équivoque concerne le caractère vague du terme : comme le terme *peine*, celui de *punition* est parfois employé de manière lâche pour désigner, d'une part, une pratique humaine relativement circonscrite et, d'autre part, un ensemble hétérogène qui recouvre des événements aussi disparates que la « punition chez les animaux » (où la punition est typiquement considérée une forme de représailles), la « punition chez les plantes et les insectes » (voir, par exemple, Nakao et Machery 2012, où la punition est pensée comme une stratégie par laquelle un organisme impose des coûts sur un autre), ou, dans un contexte religieux, « la punition du ciel » ou « la punition divine » qu'une entité surnaturelle impose sur les humains, le plus souvent pour leurs péchés.

Afin d'éviter cette double équivoque, cet article définit le terme *punition* comme une pratique distinctement humaine qui consiste à imposer une sanction sur une personne en réaction à une action ou faute que la personne a commise *et* comme une pratique qui n'a pas une visée nécessairement rétributive ou autrement afflictive. La raison du premier choix définitionnel (punition comme pratique humaine) relève d'un souci de clarté et de cohérence. La raison du deuxième choix définitionnel (punition comme sanction et non pas nécessairement comme rétribution) résulte du fait que la visée rétributive n'est qu'une motivation morale parmi d'autres susceptibles de susciter la punition. Aussi, réduire la punition à l'idée de rétribution serait réducteur.

Cet article résume et offre une analyse normative sommaire d'une série de recherches récentes sur la psychologie morale de la punition. L'accent est mis sur des études qui prennent les dimensions morales de la punition au sérieux et qui estiment que la moralité de la punition peut être étudiée empiriquement. Plus particulièrement, l'article se concentre sur les motivations pour punir que les gens ordinaires - à savoir, pas seulement les professionnels qui travaillent au sein du système pénal – révèlent dans un contexte expérimental. Quatre types de motivations morales seront envisagées dans ce qui suit : la punition altruiste, la punition rétributive, la punition préventive et la punition transformative ou communicative. La présentation de ces recherches empiriques sur chacun de ces quatre types de motivation morale sera doublée d'une brève discussion sur l'interprétation des observations empiriques.

La punition altruiste est un phénomène amplement documenté depuis les recherches d'Ernest Fehr et Simon Gächter (Fehr et Gächter 2000, 2002). La punition altruiste désigne le phénomène par lequel, dans le cadre d'un jeu expérimental, des individus choisissent de sanctionner les individus qui refusent de coopérer avec les autres même si la sanction imposée est nettement coûteuse - et, par conséquent, irrationnelle du point de vue de l'intérêt individuel - pour ceux qui l'imposent. Le jeu expérimental consiste dans un schéma où chaque participant reçoit de manière transparente une somme monétaire égale à celle des autres et où chaque participant peut choisir d'investir une fraction ou la totalité de cette somme d'une manière qui est bénéfique au niveau du groupe (en ceci que chaque investissement est au bénéficie des autres membres du groupe), mais qui n'est pas rationnellement bénéfique au niveau de l'individu (en ceci que, pour chaque investissement donné, l'individu qui investit n'a pas la garantie que les autres investiront dans un proportion égale). Il est ainsi rationnel, du point de vue de l'intérêt individuel, de ne rien investir, à savoir de ne pas coopérer avec les individus qui investissent. Fehr et

Gächter ont trouvé que, lorsqu'on donne la possibilité aux participants de sanctionner les individus qui ont choisi de ne pas coopérer, une large majorité des participants (plus de 80%) choisissent de sanctionner au moins une fois l'individu qui n'a pas coopéré. De manière significative, les participants choisissent de sanctionner la non-coopération même si les sanctions sont individuellement coûteuses, à savoir que ceux qui les imposent *savent* que, en vertu du caractère isolé et unique du jeu, ils ne pourront pas bénéficier d'un potentiel changement coopératif dans le comportement des individus sanctionnés. Puisque ces sanctions ne servent pas l'intérêt de ceux qui les imposent, Fehr et Gächter appellent ce phénomène la punition altruiste.

Le syntagme *punition altruiste* est pourtant potentiellement inadéquat. Ceci est parce que les émotions qui semblent motiver les individus à sanctionner ne sont pas des émotions spécifiquement altruistes (par exemple, l'empathie), mais, comme Fehr et Gächter le suggèrent, des émotions négatives comme la colère ou l'irritation. Néanmoins, cette explication de la punition altruiste par des émotions négatives est sujette à au moins deux limitations. Premièrement, ces émotions ne sont pas nécessairement négatives. Comme le note Jon Elster (2005), la motivation pour punir pourrait résider dans un sentiment de lueur chaude (*warm glow*), à savoir dans le sentiment de satisfaction que les individus qui punissent ressentent *par rapport à eux-mêmes* en anticipation de la sanction qu'ils imposeront sur ceux qui n'ont pas coopéré. Punir, dans ce cas, serait motivé par le plaisir que la personne qui punit prend à son image de personne qui ne tolère pas la non-coopération. Deuxièmement, il n'est pas empiriquement clair *quelles* sont précisément les émotions qui régissent la punition altruiste (l'altruisme n'étant pas, d'ailleurs, une émotion). Par exemple, l'étude de Pedersen et al. (2013) montre que ce ne sont pas tellement la colère ou l'irritation ressenties face à la transgression des normes de coopération qui motivent la punition de ceux qui ne coopèrent pas. Pour Pedersen et al. (2013), c'est plutôt la jalousie ressentie face aux gains remportés par ceux qui ne coopèrent pas qui explique la sanction qu'on leur impose.

Le deuxième type de motivation pour punir relève, au moins en partie, de l'intérêt individuel. Ainsi, il existe des explications qui posent que l'on punit pour prévenir (ou, au moins, à réduire la probabilité de) la répétition des actes et comportements punis dans le futur. Selon cette hypothèse, la punition pourrait être expliquée de manière (quasi-)évolutionniste: quelle que soit la motivation immédiate qui rend compte du fait que l'on punit, la punition est un comportement social sélectionné pour sa fonction adaptative. Cette explication avance que la punition sert l'intérêt individuel en ceci qu'elle permet de tenir les tricheurs à distance. Ainsi, il est rationnel de punir ou, plus exactement, de soutenir la pratique sociale de la punition parce que *ne pas punir* encouragerait des comportements qui contreviennent à nos intérêts individuels immédiats, comme l'intérêt dans notre intégrité corporelle ou l'intérêt à avoir notre propriété protégée. Autrement dit, la motivation pour la punition est rationnellement préventive du point de vue des intérêts individuels et explicable de manière évolutionniste de par sa fonction adaptative. Cette explication est articulée, entre autres, par Morris Hoffman (2014).

Recourir à une explication évolutionniste des motivations pour punir est problématique à plusieurs niveaux. Premièrement, même si l'explication était logiquement valide, il reste qu'il n'y a pas de données empiriques pour montrer que la décision de punir est directement causée ou causalement médiée par un calcul rationnel. Deuxièmement, tenter d'échapper au premier problème en avançant qu'il existe des émotions qui pourraient être interprétées *comme si* elles agissaient par procuration de nos intérêts individuels – en ceci que punir à cause d'une émotion était comme si l'on punissait par choix rationnel – placerait l'explication par l'intérêt rationnel dans la catégorie conceptuellement suspecte des explications purement hypothétiques, à savoir des explications qui n'arrivent pas à rendre compte des mécanismes explicatifs qu'elles ont pourtant pour tâche principale de spécifier (pour

une critique de cette catégorie d'explications, voir Elster 2015). Troisièmement, la nature fonctionnaliste de l'explication évolutionniste suppose que les causes motivationnelles de la punition résident dans ses conséquences, à savoir, la prévention des comportements contraires à nos intérêts. Ceci rompt avec la logique propre aux explications causales, où les causes précédent les effets (voir Elster 2015). Qui plus est, puisque la punition est une pratique humaine généralement (sinon universellement) répandue, il est difficile (sinon impossible) de montrer comment l'explication évolutionniste de la punition pourrait être testée empiriquement.

Le troisième type de motivation pour punir – à savoir, la rétribution – est corroboré par une série de recherches expérimentales qui permettent de rejeter l'hypothèse selon laquelle la punition serait motivée par des considérations rationnelles liées à la prévention. Par exemple, le sondage expérimental mené par Carlsmith, Darley et Robinson (2002) montrent que, bien que les individus affirment souvent que la motivation centrale pour punir *devrait* être la prévention, c'est la rétribution qui agit comme la motivation réelle lorsque les participants doivent décider quelle sanction il faut imposer aux individus qui ont commis une action moralement condamnable. Plus précisément, les individus préfèrent imposer des sanctions dont la sévérité est proportionnelle à la gravité morale des faits punis alors même qu'une sanction moins sévère peut garantir la prévention des faits punis. Bien que cette explication permette d'écarter l'hypothèse selon laquelle la prévention motiverait la punition, le fait d'identifier la motivation rétributive à un sens pour la proportionnalité est contestable. Ceci est parce que le sens pour la proportionnalité n'est pas nécessairement substantiellement rétributif. La préférence pour la proportionnalité punitive pourrait, par exemple, découler d'une préférence pour l'impartialité dans la distribution des sanctions ou d'une volonté à exprimer la reconnaissance du tort moral. De plus, la conception que Carlsmith, Darley et Robinson (2002) avancent de la rétribution ne retient pas la composante afflictive qu'on lui associe de manière intuitive, selon laquelle la rétribution consiste dans le fait de vouloir faire souffrir ou d'imposer un coût hédoniquement sévère à la personne punie.

L'hypothèse selon laquelle l'une des motivations centrales pour punir consiste à faire souffrir ou à imposer un coût à la personne punie a été testée et partiellement corroborée par Nadelhoffer et al. (2013), ainsi que par Crocket, Özdemir et Fehr (2014). Ces études expérimentales montrent que les participants sont prêts à imposer une sanction monétaire sur des personnes ayant commis des actions condamnables même si ceux qui punissent savent que les personnes punies n'auront pas conscience ou n'éprouvent pas le coût qu'on leur impose. Il semble ainsi que le désir de faire souffrir la personne punie compte parmi les motivations brutes – à savoir, irréductible à d'autres motivations – pour la punition. Toutefois, le nature économique de ces études en soulève une limitation importante, en ceci qu'elles ne permettent pas de séparer la volonté de faire souffrir la personne punie de la volonté de lui enlever un gain acquis de manière immorale.

Le quatrième et dernier type de motivation pour punir pointe vers la volonté de faire parvenir un message moral à la personne punie. Punir, dans ce sens, constituerait un acte par lequel on communique à la personne punie un message qui souligne le caractère moralement condamnable de son action avec l'intention de l'amener à reconnaître sa faute et à agir en conséquence de cette reconnaissance. Les jeux expérimentaux menés par Frederieke Funk (2015) montrent qu'il n'a pas de corrélation significative entre la perception de la souffrance subie par la personne punie et le degré dans lequel la punition est ressentie comme satisfaisante par ceux qui punissent. Par contre, le changement d'attitude de la personne punie – par exemple, le fait de comprendre la faute commise et de demander pardon – est significativement corrélé à la satisfaction ressentie par ceux qui punissent. Pour Funk, ceci indique que les motifs purement rétributif ne jouent pas un rôle motivationnellement central dans la punition : si la motivation pour punir était

strictement le fait de faire souffrir les personnes punies, il y aurait une relation significative entre la souffrance perçue de ces dernières et la satisfaction ressentie par ceux qui punissent. La motivation pour punir réside dans la volonté de communiquer un message moral à la personne punie et de susciter ainsi un changement dans son attitude face à l'action punie.

L'une des limites principales de l'interprétation de Funk est que l'on ne peut pas, à strictement parler, inférer la nature précise de la motivation pour punir à partir de la satisfaction ressentie après le moment de la punition. Ceci est parce que la satisfaction morale ressentie face à la punition pourrait être influencée par des considérations qui ne sont réellement intervenues qu'une fois que la punition a été imposée. Aussi, Funk n'a pas exclu la possibilité que le désir de voir la personne punie comprendre le message moral de la punition soit motivationnellement absent au moment où l'on décide de punir.

Ce sommaire des différentes recherches sur la psychologie (morale) de la punition permettent de déceler au moins deux conclusions préliminaires. La première conclusion est que punition est motivationnellemenent surdéterminée, à savoir qu'elle n'est pas déterminé par un seul type de motivation (morale). La deuxième conclusion est que l'intérêt individuel joue un rôle minimal, sinon négligeable, dans l'explication des motivations pour punir.

Bibliographie

CARLSMITH K.M., DARLEY J.M., ROBINSON P.H. «Why do we punish ? Deterrence and just deserts as motives for punishment », *Journal of Personality and Social Psychology*, 2002, vol. 83, n° 2, p. 284-299

CROCKET M.J., ÖZDEMIR Y., FEHR E. « The value of vengeance and the demand for deterrence », *Journal of Experimental Psychology*, vol. 143, n° 6, p. 2279-2286

ELSTER J. «Fehr on Altruism, Emotion, and Norms », *Analyse & Kritik*, 2005, vol. 27, p. 197-211

ELSTER J. *Explaining Social Behavior : More Nuts and Bolts for the Social Sciences*, 2e éd., 2015, Cambridge University Press

FEHR E., GÄCHTER S. « Cooperation and punishment in public goods experiments », *American Economic Review*, 2000, vol. 90, p. 980-984 ;

FEHR E., GÄCHTER S.. « Altruistic punishment in humans », *Nature*, 2002, vol. 415, p. 137-140.

HOFFMAN, M. *The Punisher's Brain : The Evolution of Judge and Jury*, Cambridge University Press, 2014

NADELHOFFER T., HESHMATI S., KAPLAN D., NICHOLS S. « Folk Retributivism and the Communication Confound », *Economics and Philosophy*, 2013, vol. 29, p. 235-261

NAKAO H., MACHERY, E. « The evolution of punishment », *Biology and Philosophy*, 2012, vol. 27, n° 6, p. 833-850

PEDERSEN E.J., KURZBAN R., MCCULLOUG M.E. «Do humans really punish altruistically? », *Proceedings of the Royal Society B*, 2013, vol. 280, n° 1758, p. 20122723