Review

Common liability to addiction and “gateway hypothesis”: Theoretical, empirical and evolutionary perspective

Michael M. Vanyukov\textsuperscript{a,b,c,*}, Ralph E. Tarter\textsuperscript{a,b}, Galina P. Kirillova\textsuperscript{a}, Levent Kirisci\textsuperscript{a}, Maureen D. Reynolds\textsuperscript{a}, Mary Jeanne Kreek\textsuperscript{d}, Kevin P. Conway\textsuperscript{e}, William G. Iacono\textsuperscript{g}, Laura Bierut\textsuperscript{b}, Michael C. Neale\textsuperscript{i}, Duncan B. Clark\textsuperscript{b}, Ty A. Ridenour\textsuperscript{a}

\textsuperscript{a} Department of Pharmaceutical Sciences, University of Pittsburgh, Pittsburgh, PA, USA
\textsuperscript{b} Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA
\textsuperscript{c} Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA
\textsuperscript{d} Laboratory on the Biology of Addictive Diseases, Rockefeller University, New York, NY, USA
\textsuperscript{e} Division of Epidemiology, Services and Prevention Research at the National Institute on Drug Abuse (NIDA), Bethesda, MD, USA
\textsuperscript{f} Department of Mental Health, Johns Hopkins University, Baltimore, MD, USA
\textsuperscript{g} Department of Psychology, University of Minnesota, Minneapolis, MN, USA
\textsuperscript{h} Department of Psychiatry, Washington University, St. Louis, MO, USA
\textsuperscript{i} Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA

A R T I C L E   I N F O

Article history:
Received 6 September 2011
Received in revised form 12 December 2011
Accepted 13 December 2011
Available online xxx

Keywords:
Drug dependence
Drug abuse
Genetics
Phenotype
Evolution

A B S T R A C T

Background: Two competing concepts address the development of involvement with psychoactive substances: the "gateway hypothesis" (GH) and common liability to addiction (CLA).
Method: The literature on theoretical foundations and empirical findings related to both concepts is reviewed.
Results: The data suggest that drug use initiation sequencing, the core GH element, is variable and opportunistic rather than uniform and developmentally deterministic. The association between risks for use of different substances, if any, can be more readily explained by common underpinnings than by specific staging. In contrast, the CLA concept is grounded in genetic theory and supported by data identifying common sources of variation in the risk for specific addictions. This commonality has identifiable neurobiological substrates and plausible evolutionary explanations.
Conclusions: Whereas the “gateway” hypothesis does not specify mechanistic connections between “stages”, and does not extend to the risks for additions, the concept of common liability to addictions incorporates sequencing of drug use initiation as well as extends to related addictions and their severity, provides a parsimonious explanation of substance use and addiction co-occurrence, and establishes a theoretical and empirical foundation to research in etiology, quantitative risk and severity measurement, as well as targeted non-drug-specific prevention and early intervention.

© 2012 Elsevier Ireland Ltd. All rights reserved.


* Corresponding author at: University of Pittsburgh, Department of Pharmaceutical Sciences, 3520 Forbes Avenue, Suite 203, Pittsburgh, PA 15213, USA.
E-mail address: mmv@pitt.edu (M.M. Vanyukov).
1. Introduction

Substance use disorders (SUD) comprise a clinically heterogeneous group of conditions. In part this heterogeneity is due to difficulties in defining the disorder consequent to drug use. Apart from the involvement of criteria other than medical (e.g., legal, cultural) in defining the normative drug-pertaining behavior, the SUD phenotype has not been uniformly described, and there has been disagreement as to what it should be labeled, e.g., dependence or addiction (O’Brien et al., 2006). Whereas dependence is accepted in the current classification (DSM, ICD), as based on physiological drug response, addiction, defined as compulsive drug-seeking and use, may be more relevant to the designation of the overall clinically important phenotype. Dependence is a normal physiological adaptation to drug action, which may occur outside of drug abuse context (e.g., during pain treatment): it is addiction, a pernicious behavior, that results in health and other ensuing problems. This behavior frequently, but not necessarily, results from or is contributed by dependence, whereas dependence does not necessarily manifest in addictive behavior.

Human behaviors and their deviations often have social significance, which influences definition of deviant forms. In particular, when such significance is codified in a legal system, as is the division into licit and illicit substance use, the terminology is frequently value-laden and fraught with stigmatization. This is illustrated, for instance, by a discussion of choice between “addiction” and “dependence” (Erickson and Wilcox, 2006; O’Brien et al., 2006), in which the proponents of each term refer to stigmatization as a main shortcoming of the alternative label. It is, however, hardly possible to find a term for an illicit behavior that is independent of societal norms and devoid of emotional content. Any newly introduced label will likely acquire the same negative connotations as the currently used ones, because illicit behaviors evoke negative reactions from a large proportion of the population. Even though addiction is a medical/psychiatric disorder, it results from voluntary norm-violating behaviors and thus is subject to moral judgment regardless of the term used. Whereas an addict may have lost some control over his behavior, becoming slave to a habit, this slavery is viewed as voluntarily self-inflicted and thus reprehensible. Appropriately, addictus was a term applied in ancient Rome to a person in a legal slave-like condition (for delinquent debt, which, if left unpaid by an addictus in 30 days, would result in slavery and possibly death) (Smith, 1898).

Whereas the social and legal status of a drug is an environmental factor, it profoundly influences causes of individual behavioral variation pertaining to its use, including genetic sources of this variation. It is the legal definition of substances as licit/illicit, reflecting social conventions and significance of related behaviors rather than drug properties or organismic reactions to drugs, that results in a corresponding pattern of genetic clustering of liabilities to the respective disorders, forming two genetically distinct albeit correlated groupings of dependence symptoms respectively for licit and illicit drugs (Kendler et al., 2007). It is thus likely that a change in a drug’s legal status (such as those, e.g., that happened with alcohol related to Prohibition) would result in the respective addiction’s migration to the genetically and etiologically different group.

The heterogeneity of SUD/addictions is also partly due to the drugs’ differences in chemical classes, routes of administration, neurobiological systems they act on, metabolic pathways, and psychopharmacologic effects. Accordingly, SUD are classified into groups defined by chemical origins of the drug or its effects. This classification, however, does not resolve clinical heterogeneity, as the categorical substance-specific SUD diagnosis corresponds to hundreds of different combinations of symptoms (Vanyukov et al., 2003b). This vagueness of the diagnostic threshold creates vast variation within groups even with the same drug-specific diagnosis.

Taking into account this heterogeneity of the clinical phenotype, it may seem surprising that the risks for addictions related to the specific drugs are highly correlated. Considering that these risks have substantial (moderate to high) heritability, i.e., genetic contribution to phenotypic variance, it is indeed noteworthy that this contribution is largely non-specific, at least for addictions to licit drugs (Kendler et al., 2003). Therefore, variation in the substance-specific addiction risks is largely due to shared biological mechanisms, leading to compulsive drug seeking, a defining feature of addiction (Conway et al., 2010; Koob and Volkow, 2010). This commonality is thus unlikely to derive from drug metabolism.

Two competing frameworks have been proposed to explain the origins of this commonality and account for the development of drug involvement and co-occurrence of addictions to different drugs. The “gateway hypothesis” (GH) connects initiation of use of various drugs by a sequence stated to represent a developmentally staged process. The alternative model of common (general) liability to addiction (CLA) posits non-specific liability to all drug addictions, regardless of the order of use initiation. The two perspectives have substantially different ramifications for research, clinical practice, and health policy.

2. Gateway hypothesis

Research concerned with the GH, during the 35 years since it was advanced (Kandel, 1975), has commanded substantial attention due to the high practical value that has been ascribed to the observation of a sequential order in drug use initiation. GH was predated by the similar “stepping-stone” theory that first appeared in the 1930s and assumed that consumption of a “soft” drug such as marijuana inexorably sets an individual on a trajectory to addiction to hard drugs. GH relaxes the inevitability assumption, but still posits that substance use starting with a licit substance and progressing to marijuana leads to use of other, “harder” drugs. Marijuana is thus designated the “gateway” drug in this progression, although originally, and sometimes still, this role has been assigned to alcohol and tobacco (Grunberg and Faraday, 2002; Kandel, 1975). Because it is frequently observed that “[v]ery few individuals who have tried cocaine and heroin have not already used marijuana; the majority have previously used alcohol or tobacco” (Kandel, 2003, p. 482), the conclusion is drawn (Kandel, 2002a) that “there is a progressive and hierarchical sequence of stages of drug use that begins
with tobacco or alcohol, two classes of drugs that are legal, and proceeds to marijuana, and from marijuana to other illicit drugs, such as cocaine, metamphetamines [sic], and heroin" (p. 3). The sequence of drug use initiation is thus the essence of the GH. Accordingly, it is drug use itself that is viewed as the cause of drug use development. Likewise, the “stages” are defined in a circular manner: a stage is said to be reached when a certain drug(s) is used, but this drug is supposed to be used only upon reaching this stage. In other words, the stage both is identified by the drug and identifies that drug. In effect, the drug is identical to the stage (“marihuana is a crucial stage . . .”) (Kandel, 1975). There is no process or organismic characteristic, separate from drug use per se, which is presumed to cause or underlie the supposed developmental staging indicated by, and identical to, drug milestones. Moreover, the notion of “stage” itself is redefined from its common meaning, such that the later stages are assumed to be reachable, albeit less frequently, before the earlier ones. This renders the GH incongruent with the conventional biological developmental framework (cf. stages of embryonic development).

Proponents of the GH avoid explicit assertions regarding causality (“causal claims in the Gateway Hypothesis . . . are still beyond reach” (Kandel and Jessor, 2002, p. 371)). Nonetheless, causation is implicit in the statements such as “one illicit drug is required [emphasis added] to make the progression to marijuana use” (Kandel and Yamaguchi, 2002, p. 71), or “[t]he use of a drug at a lower stage is necessary [emphasis added] but not sufficient for progression to a higher stage indicating involvement with more serious drugs” (Kandel and Yamaguchi, 2002, p. 69). It is also stated that the validity of the GH is based on (1) the sequence of drug use initiation, and (2) “association in the use of drugs, such that use of a drug(s) lower in the sequence increases the risk of using drugs higher in the sequence. Ultimately, association implies causation if all possibilities for spurious associations have been eliminated.” It is only because of “the difficulties of establishing true causality in the social sciences” that “the term association rather than causation is emphasized . . .” (Kandel, 2002a, p. 4). Causality is, however, readily refuted by the frequently observed “atypical” sequencing (see below) as well as the lack of “true association”. Even if present, association does not need to be spurious to be non-causal. A common “cause”, source of variance in the risk, may be a plausible explanation of sequential use, particularly when one drug is more available than another.

Apart from these refutable assertions, it is hard to discern a falsifiable, i.e., hypothetical, element in the GH beyond the sequence observation. Virtually every proposition of the GH is qualified by a disclaimer, effectively engulfing and neutralizing possible arguments to the contrary. Indeed even the core facet of the GH is hedged: “[t]he notion of developmental stages in drug behavior does not imply, however, that these stages are either obligatory or universal, nor that all persons must progress through each in turn” (Kandel, 2002a, p. 3) (cf. the above quoted contradicting statements of necessity to pass a “lower stage” and the requirement for using of a licit drug to progress to marijuana). After decades of research, even “[t]he notion of a Gateway drug itself is vague” (p. 7).

A brief look at the sequence further belies the significance of the postulated “stages”. (1) Indeed, unsurprisingly, the use of any substances is preceded by non-use (the only truly invariant step of the sequence). It is the non-use then, which, by the logic of the “majority have previously used . . .” line of thinking, should be the actual gateway condition. (2) The first use of illicit drugs is frequently preceded by consumption of licit substances. Following the logic of the current implementation of the gateway theory, it is then alcohol and tobacco use rather than marijuana and “hard” drugs that need to be prevented, because once the licit outset of the sequence is barred, there should be no danger that the rest of it will materialize. Common sense and experience, however, suggest otherwise. (3) The initiation order is frequently reversed even for the licit-to-illicit sequence, in contradiction to the stated first validity criterion of the GH. This order becomes even less consistent beyond involvement with illicit substances, to the degree that the use of illicit drugs other than marijuana, i.e., “hard” drugs, is usually collapsed in one class in GH modeling use sequences (Kandel and Yamaguchi, 1999). The use of “hard” drugs, such as cocaine and heroin, is frequently preceded by “soft” drugs such as marijuana, whose legal status depends on the population. In fact, however, when the frequency of marijuana use in the population is taken into account, the true association between marijuana and the “hard” drug use is negligible (Earleywine, 2002). This negates the above quoted second criterion of validity for GH. The high correlations encountered in the literature and establishing the association between marijuana use and other drug use are artificial, because they are estimated among hard drug users, without taking into account the large population of those who try or even habitually use marijuana but never transition to harder drugs (Earleywine, 2002).

The reversals of the “gateway” sequence are particularly noteworthy. It is an empirical fact that a substantial proportion of drug users initiate their drug involvement with illicit rather than licit drugs or use “hard” drugs before marijuana (Golub and Johnson, 2002; Kandel and Yamaguchi, 2002; Mackesy-Amiti et al., 1997; Tarter et al., 2006). When non-US samples are taken into account, as in the large multi-national study by Degenhardt et al. (2010), the “gateway” sequence and its “violations” can be even more clearly seen as functions of the frequency of cannabis use in the population. For instance, in Japan, where cannabis is used by only 4.5% of the 18–29 year old population, while use of other illicit drugs is 4.8%, cannabis is not used first by a staggering 83.2% of the users of other illicit drugs, “violators” of the “gateway” sequence. The overall level of illicit drug use does not depend on access to the purported gateway substances, whether alcohol/tobacco or cannabis, as would be expected if the GH were true. The “gateway” role of alcohol is also refuted by the evidence that whereas the aldehyde dehydrogenase deficiency is related to lower rates of drinking, it does not predict lower rates of non-alcohol substance use (Irons et al., 2007). Early protection against, or delays in, tobacco use onset also do not result in reduced risk of involvement with cannabis or other illegal drugs (Furr-Holden et al., 2004).

Nevertheless, instead of serving as indisputable grounds to reject the GH, the numerous deviations from the “gateway” sequence are dismissed by the proponents of the GH as error or random or non-systematic patterns (Kandel, 1975; Kandel and Yamaguchi, 2002). The inadequacy of the GH in its explaining away the “errors” may have reached its high point when the gateway sequence is reduced to the gateway drug, marijuana, which has at last been proposed as gateway to licit substance use, tobacco smoking, as well (Tullis et al., 2003). The reversals of the basic GH sequence and its parallelism with ranking of the prevalences of use of respective substance categories also contradict the GH’s “premise . . . that involvement in various classes of drugs is not opportunistic but follows definite pathways” (Kandel, 2002a, p. 3).

It should also be noted that the proposed sequence applies only to the use of different drugs rather than different levels or extent of drug involvement (from use to dependence) (Kandel and Jessor, 2002) and does not extend to SUD development. Therefore, interestingly, the area of GH application is outside of the medical realm and has only a temporal connection with addiction. These limitations notwithstanding, the GH has significantly influenced policy formation (Leschner, 2002), intervention (Manski et al., 2001), and research (Kandel and Yamaguchi, 1999; Kandel, 2002b). Research, whose purpose is to inform both intervention and policy, may be hindered or misdirected if a concept lacking substance, validity and utility is accorded prominence. In turn, the targets for policy and intervention may be shifted or insufficiently focused to produce an optimal impact.

3. Common liability to addiction

3.1. Common addiction liability as a trait

In contrast to the GH, which addresses only the order of drug use initiation, the concept of common (general) liability to addiction in drug use disorders (SUD) (CLA) involves mechanisms and biobehavioral characteristics that pertain to the entire course of development of the disorder and changes in the risk. The CLA concept also overlaps with the psychological and psychopathological constructs that have been previously used to explicate addiction and its mechanisms. Liability denotes a latent (unobservable) quantitative trait that, when measured, “would give us a graded scale of the degree of affectedness or of normality” (Falconer, 1965, p. 52). These two latter broad phenotypic categories are divided by a latent threshold on the liability axis. An individual’s quantitative liability phenotype at any time point, above or below the threshold, represents a value within the norm of reaction of the genotype (Dobzhansky, 1951), the genetically determined distribution of all phenotypic values for a trait in an individual for all possible environmental conditions. A probability distribution that can be considered at each time point as well as within the entire developmental trajectory, the norm of reaction fluctuates across time in accordance with the ontogenesis of liability, the changing propensity to (risk for) or severity of the disorder.

Theoretically, the liability distribution in the population may range from the individual norms of reaction that do not include the disorder (at least, in the present range of environmental conditions) and thus totally resistant (resilient)—up to the phenotypes corresponding to the most severe fastest-developing disorder. The “gradations of normality” (the subthreshold liability phenotypes) correspond to variation in the risk (propensity), whereas “gradations of affectedness” (the suprathreshold phenotypes, which are likely to be assigned a clinical diagnosis) correspond to variation in severity, comprising the two parts of the liability distribution relative to the threshold. Applied to addiction (Conway et al., 2010; Vanyukov et al., 2003a), severity refers to the degree of maladaptive compulsive drug-seeking and using behavior displayed by an individual, and corresponds to variation in liability above the diagnostic threshold. Propensity refers to the probability of the disorder onset and corresponds to liability variation below the diagnostic threshold. Variation in propensity may manifest as psychological/behavioral precursors of addiction. These precursors have been conceptualized as problem behavior (Jessor and Jessor, 1977), overlapping with the CLA (Vanyukov et al., 1994, 1996; Vanyukov and Tarter, 2000). This particularly concerns propensity mechanisms preceding drug involvement as well as its changes—the domain pertaining to the focus of the GH. The GH perspective has been presented as a “fundamental theoretical antithesis” (Kandel and Yamaguchi, 1999, p. 68) to this concept. Such a juxtaposition is not, however, fully adequate, as the problem behavior and liability theories address mechanisms of addiction development, whereas the GH at best only describes a pattern of drug use initiation.

It should be noted that there is no contradiction between the concept of liability as a single unidimensional trait and that this trait is complex, i.e., contributed to by many diverse influences. In fact, these are common features of complex (multifactorial) traits, such as stature (observed trait) or IQ (index of the latent trait of intelligence). Normal distributions of these traits in the population result exactly from the fact that they are influenced by multiple variables, in accordance with the central limit theorem. Unidimensionality of the trait refers to the structure of covariation between these variables rather than to the total number of the manifold influences potentially determining the shared variance, or to the presence of a single causal influence. The number of such influences for CLA may be substantial. To a considerable degree, they appear to pertain to an identifiable circumscribed group of mechanisms underlying behavior regulation and socialization.

3.2. Mechanisms of variation in CLA

3.2.1. Empirical support for the common addiction liability concept. Drug- or drug-class-specific mechanisms of metabolism and pharmacologic action notwithstanding, a wealth of evidence indicates substantial commonality among the different substance use disorder categories. This commonality mitigates the possibility that the addictions each represent a discrete disorder. Both the plausibility of a single common (non-drug specific) liability dimension and the feasibility of measurement of this latent trait are supported by clinical, neurobiological, genetic and statistical findings (Vanyukov et al., 2009, 2003a,b). Thus, statistical modeling suggests that correlations of marijuana and other drug use are accounted for by common liability to and opportunities for consumption (Morral et al., 2002). Further support for the CLA concept is provided by the findings showing location of diagnoses and symptoms of SUD related to different drugs on the same dimension (Agrawal et al., 2004; Kendler et al., 2007; Kirisci et al., 2006, 2002; Wu et al., 2009).

Key support for the CLA concept comes from the high genetic correlations between liabilities to specific drug addictions (correlations between the genetic components of the liability variance) determined in biometrical genetic studies. Virtually no specific genetic variance is estimated (Kendler et al., 2003), aside from the licit/illicit groupings related to the legality of use (Kendler et al., 2007), i.e., to the behavioral response to societal norms. Non-drug-specific mechanisms, e.g., reflected in neurobiological data pertaining to drug-related reinforcement, suggest commonality of many drug effects involving dopaminergic and other major neurobiological systems, despite differences in the routes of administration, transformation pathways, and primary targets of psychoactive substances. Importantly, these systems substantially overlap with those that are involved in the mechanisms of behavior regulation, natural reward and incentive motivation, stress response, and social behavior (rev. in Vanyukov et al., 2003b).

3.2.2. Ontogenesis of the liability phenotype. While there is no single “substance abuse personality”, a certain proportion of variation in the liability to SUD is shared in common with personality/behavioral phenotypic variation predating substance use initiation. Initiation, however, is a necessary but not sufficient condition for addiction development. Individual phenotypic characteristics prevail over access to drugs in the determination of variation in propensity to addiction, particularly taking into account phenotype (and genotype)–environment correlations, whereby individuals with certain characteristics are more likely to attain environment facilitating access to drugs (Kirilova et al., 2008). This likelihood further grows once drug use commences and becomes habitual, regardless of the drug used. In other words, the ontogenesis of the addiction liability phenotype further strengthens the effect of the genotype and of the initial/prior phenotype due to compounding of the phenotype–environment correlation over time. Growth in heritability with age (Hicks et al., 2007) is likely caused by that compounding.

The role of the individual characteristics (and of the genotype) in the variation of addiction risk is particularly eminent and grows with age also due to the ready availability of psychoactive substances. For instance, according to the 2007 National Survey of Drug Use and Health (U.S. Dept. of Health and Human Services, 2008), half of the population aged 18–49 have used marijuana, 70% have used cigarettes, and 90% have used alcohol. Almost a quarter of the population aged 35–49 have used cocaine. Variation in liability to drug use initiation is substantially contributed by shared and nonshared environment (Han et al., 1999; Kendler et al., 2000), in contrast to...
liability to SUD (abuse and dependence) where the genetic component of variance is predominant and no shared environment effect is detected. This difference suggests that drug use initiation, while obviously a precondition to addiction, does not account for the risk for SUD. Supporting the role of phenotypic propensity in the risk for addiction, the transmissible liability index, TLI (Vanyukov et al., 2009, 2003a,b), a scale based on psychological indicators of CLA prior to exposure to drugs, was higher in boys who later developed cannabis use disorder compared to those who used cannabis but did not develop the disorder (Kirisci et al., 2009). It is noteworthy that no shared environment component (non-genetic sources of twins’ similarity) was detected for TLI in a twin study (Vanyukov et al., 2009). SUD or substance use has been used as an indicator of the highly heritable latent trait of disinhibition/externalizing (Clark et al., 2008; Iacono et al., 2008; Krueger et al., 2002; Young et al., 2000), supporting the role of behavior regulation/disinhibition as a developmental component of SUD liability (Tarter et al., 1999, 2003). Disinhibition, as indicated by hyperactivity, impulsiveness, antisociality/psychopathy, has long been hypothesized to be genetically related to substance abuse (e.g., Gorenstein and Newman, 1980).

Unlike simple monogenic and some complex traits that are relatively static (e.g., stature in adults), liability to SUD is dynamic, likely characterized by a nonlinear developmental trajectory (Tarter and Vanyukov, 1994). As shown in Fig. 1, concrete organicistic and environmental factors can be conceptualized as vectors projecting onto the liability axis, analogous to vectors of force in physics. These vectors determine the direction and position of the trajectory at any point in time, with the initial location of the individual phenotype on the liability axis at the moment of conception. Considered as a vector, even a small relevant functional genetic or early environmental difference, being at the origin of a causative chain as it is, gives an individual behavioral phenotype the initial push that may result in a large deviation in later development, accumulating momentum. At the relatively distal time when genetic relationships are usually evaluated, i.e., when a disorder or a behavioral deviation can be observed, the role of this early-acting factor may be as difficult to detect as it is to find a match that started a forest fire. The relative rarity and problems with the reproducibility of positive findings in molecular genetic studies of behavioral traits are thus not surprising. The complex and interacting genetic mechanisms are at the very outset of the long developmental process, and their immediate contribution is modulated by epigenetic modifications and other factors influencing gene expression.

The same factor may produce opposite effects in different individuals. For instance, parental substance abuse as an environmental factor may both promote similar behavior in offspring (by providing a ready access to alcohol and other drugs) and produce aversion to it (by demonstrating clearly the negative effects). The vectors may interact; e.g., the role of genetic factors may be contingent on environmental ones, and vice versa, as, for instance, was shown for heritability of disinhibition (a scale evaluating the need to disinhibit behavior in the social sphere by drinking,party-ing and seeking variety in sexual partners), which depended on receiving a religious upbringing (Boomsma et al., 1999). Particular genes or environmental factors may be of special importance for an individual trajectory regardless of heritability, and an individual trajectory may have little to do with the forces detectable at the population level (an example of the strong association between a rare mutation in the MAOA gene and aggressive behavior (Brunner et al., 1993) illustrates this point). The sets of vectors may also undergo secular changes and vary across different populations.

The malleability of the liability phenotype, including “maturing-out” of conditions satisfying diagnostic criteria for dependence (e.g., Dawson et al., 2006), holds promise for effective intervention. To potentially guide research, the time points for the ontogenetic trajectory can be specified as age periods, at which distinct factors influencing addiction liability and its development may be identified. For instance, the genotype, which is the program of this development with all its individually possible variants, is determined at the moment of conception. The range of these variants, comprising the individual norm of reaction, is narrowed during development as the liability phenotype forms under individual environmental conditions. Some of those environmental influences are profound (e.g., intrauterine exposure to teratogens impairing brain development), may have various lengths of period-specificity related to the different organizational levels of impact and different developmental periods (e.g., lead neurotoxicity, attachment problems, maltreatment, imitation of “adult” behavior), and are likely mediated by epigenetic changes. A major growth in risk, corresponding to a shift of the population liability distribution to the right, occurs during the perinatal period and is strongly associated with disruptive behavior disorders. This association has been theorized by us and others to result from the common mechanisms related to deviations in neurologic maturation (Bauer and Hesselbrock, 2003; Mezzacappa et al., 1999; Tarter et al., 1999).

The structure of genetic correlations between liability to addiction and other behavioral traits indicates that sources of CLA are also likely to be unrelated to specific drug action as such, including a specific role of a drug designated as a “gateway.” It is possible, for instance, that marijuana use leads to sensitization, increasing the response of neurobiological systems involved in addiction to other drugs. Sensitization has been proposed as “probably the most relevant interpretation underlying the Gateway Hypothesis” (Kandel et al., 2006, p. 471). Apart from the data suggesting the lack of marijuana sensitization (Ellgren et al., 2004), however, a possibility of sensitization clearly exists for any other psychoactive drug. Therefore, it would not support the specific “gateway” role of marijuana as the initiator of the sequence and thus does not concern the GH, or it would support any drug’s “gateway” role, rendering the concept superfluous. Moreover, the sensitization mechanisms would have to be by definition shared in common between marijuana and other drugs.

The difference between the nonsensitized and sensitized states may be effectively the same as between individuals whose liability phenotypes differ: the threshold for use, contingent on social norms and perceived hedonic effects, is likely to be lower in an

individual with high liability. In fact, if sensitization does occur, it is equivalent to an increase in individual liability at the level of neurochemical mechanisms of addiction. This is but one way by which the labile individual liability phenotype may fluctuate, as it does, for instance, when access to alcohol or drugs is attained or precluded, when peer pressure to use alcohol or drugs is experienced or ceases, or when stressful conditions motivate self-medication alcohol or drug use. The perceived hedonic benefit from drug use due to sensitization may or may not be a significant factor furthering drug use, wherein it may be combined with augmented access to drugs, decreased perception of danger and societal disapproval, and indeed with ignoring or approval of this behavior as normative—in certain social groups. Regardless of the role of sensitization, the probability of drug use initiation with a certain drug corresponds to its frequency of use in the population and availability. In the CLA theory, sequencing of drug use is thus not denied but is viewed as opportunistic and trivial.

Data show that drug dependence can indeed be located on the same dimension as premorbid (and even pre-drug-use) behaviors that are indicators of a highly heritable latent trait variably referred to as dysregulation, disinhibition, behavior undercontrol or externalizing behavior, including risks for disruptive behavior disorders (Button et al., 2007; Iacono et al., 2008; Krueger et al., 2002, 2007; Tarter et al., 1999, 2003). At least to some degree, the association of antisocial behavior and illicit drug use is due to the illegality and thus antisociality of drug use. This trait, therefore, can be viewed as substantially overlapping with CLA (Iacono et al., 2008; Vanyukov et al., 2003b). From this perspective, drug use and related addiction are facets of a behavioral set involved with societal norms and their violations. Considering the heritability of this behavioral set, variation in the ability to assimilate these norms is to a large degree based on the individual rather than contextual characteristics.

3.3. Sources of common variance

Such individual characteristics – potential sources of common variance between liabilities to specific addictions as well as between these liabilities and other behavioral traits – have been recorded at all levels of biological organization. In particular, psychophysiological correlates of disruptive behavior disorders (attention deficit hyperactivity and conduct disorders) and SUD are consistent with deviations in frontal and prefrontal brain maturation (Bauer and Hesselbrock, 2003). Importantly, these brain areas are involved in both behavior regulation and reward, including drug-related reinforcement. The density of dopaminergic, adrenergic, serotonergic, cholinergic, and GABAAergic receptors, all involved in response to psychoactive drugs, parallels synapticogenesis (Lidow et al., 1991). Overproduction and subsequent pruning of synaptic contacts are characteristic of the peripubertal period, when the risk for non-normative behavior is maximized. Pubertal maturation rate influences neurodevelopment, with these ontogenetic changes reflected in mental processes germane to cognition, affect, and behavior regulation (Teicher et al., 2003), and thus to the ontogenesis of the SUD liability phenotype. Variation (including sex differences) in the rate of neuronal pruning in adolescence, particularly dopamine neurons in striatum, has behavioral and stress reactivity consequences; in turn, early stress may produce precocious maturation of the prefrontal cortex, leading to its underdevelopment (Teicher et al., 2002). This process and addiction itself are manifestations of brain plasticity that appear to operate with the same neuroanatomical mechanisms and systems. Notably, parental SUD has a positive dose–effect relationship with the rate of pubertal maturation in boys, related in turn with their behavioral dysregulation (Kirillova et al., 2001, 2008). Inasmuch as the familiarity of both SUD liability and behavioral dysregulation is largely due to heritability rather than common environment, these data suggest that genetic mechanisms of neural maturation and plasticity are involved in the transmission of addiction liability.

These mechanisms include an intricate and intensively studied system of relationships involving stress- and socialization-related hormones (including cortisol, oxytocin and vasopressin), behavior and affect, and neurotransmission during maturation and adolescence (Cameron, 2004; Insel, 2003; Light et al., 2004; Walker et al., 2004), manifesting sex dimorphism in many of its components. Belsky et al. (1991), based on evolutionary considerations, hypothesized that the child’s perceived lack of resources will lead to the development of “behavior patterns that function to reduce the age of biological maturation” (p. 650), whereas the opposite effect will result from the availability of resources. This theory does not exclude the possibility of an alternative, genetic, explanation of variation in developmental timing. Although the data confirm the relationship between parental absence and a lower age of menarche (Moffitt et al., 1992), it is not certain that it is due to resource-related stress, and the presence of the mother’s new partner may be of greater relevance than the father’s absence (Ellis and Garber, 2000). Early maturation in girls and boys has been shown to be associated with delinquency, mediated by peer deviance (Caspi et al., 1993: Kirillova et al., 2008), as well as higher level of distress (Ge et al., 1996), family adversity and anxiety (Tremblay and Frigon, 2005), involving conduct and antisocial personality disorders (Graber et al., 2004) and substance use and abuse (Stice et al., 2001). Sex hormone and other sex dimorphic mechanisms are not mutually exclusive, and interactions may be expected between the processes (and underlying genes) involved.

It has been noted that “virtually any manipulation that produces an enduring change in behavior leaves an anatomical footprint in the brain” (Kolb et al., 2003, pp. 3–4). In the case of drug use, this manipulation involves the systems and neural networks that may developmentally be already suboptimally prepared for effective psychological regulation. It is well known that intake of virtually all drugs activates the dopaminergic system (DS), which supports its role in the origins of common addiction liability and its variation (Vanyukov et al., 2003b). Whereas specific drugs enter the common circuit at different points, the shared structures include neurons of the ventral tegmental area (VTA), which are connected to the basal forebrain (the nucleus accumbens, olfactory tubercle, amygdala, and frontal and limbic cortices), and opioid peptide neurons within these connections (Koob and Le Moal, 2001). The DS therefore is involved with mechanisms acting both pre- and post-onset of drug use.

To be sure, numerous other systems, from the main excitatory and inhibitory, glutamate and GABA, to virtually all neurobiological systems, are likely to be included as well. Afferents of both glutamate and GABA systems modulate the activity of dopaminergic neurons in the VTA (Giogetti et al., 2002), and metabotropic glutamate receptors interact with dopaminergic neurotransmission in the nucleus accumbens (rev. in Breyssse et al., 2002)—both regions integral to the dopaminergic circuitry. Stimulation of the GABAA receptors leads to an increase in dopaminergic activity, whereas activation of GABAB receptors results in its decrease (Goudreau et al., 1994; Kalivas et al., 1990). In addition to their role in behavior regulation, the GABA- and glutamatergic systems massively contribute to drug response. For instance, the activation of serotonergic axons in the hypothalamus causes the release of metenkephalin in the VTA, which suppresses the release of GABA. This results in dopamine release in the nucleus accumbens and hippocampus. Opioid receptor agonists increase extracellular dopamine levels within the nucleus accumbens by disinhibiting GABA interneurons in the VTA (Johnson and North, 1992). Benzodiazepines and barbiturates affect GABA receptor function, and may also act in the GABAAergic circuitry that is efferent to the mesolimbic DS (Wise, 1998). Benzodiazepines increase stimulation of certain GABA
receptors in inhibitory interneurons in the VTA, which increases dopaminergic VTA excitability (Tan et al., 2010). Mice without the metabotropic glutamate receptor mGlur5 are unresponsive to cocaine (Nestler, 2001). Glutamate (NMDA) receptor function depends on the phosphorylation of CREB, a transcription factor (Das et al., 1997; Konradi et al., 1996), which is mediated by the D1 dopamine receptors. Sensitization by drugs of abuse may be facilitated by a glutamate–dopamine interaction caused when drugs are administered in a novel environment (Uslaner et al., 2001). Glutamate, via induction of the AMPA glutamate receptor Glur2 subunit in the nucleus accumbens, appears to mediate the function of another transcription factor, ΔFosB (Kelz et al., 1999). The latter has been termed a “molecular switch for addiction,” as it parallels sensitization to drugs (Nestler, 2001), and is generally an important component of neuronal plasticity (McClung et al., 2004) involved in addiction pathogenesis. Importantly, CREB inhibits drug reward, whereas ΔFosB enhances it (Nestler, 2001). Under non-novelty conditions amphetamine induces c-Fos in striatal neurons positive for dopamine D1 receptor mRNA; under conditions of novelty c-Fos is induced in both D1 and D2 neurons (Badiani et al., 1999).

The listed mechanistic elements certainly do not exhaust all potential sources of SUD liability variance even at the neurobiological level. Depending on the focus and the biological level of research, addiction has been characterized as a disorder of the brain, learning, memory, neumoraturation and neuroplasticity, homeostatic regulation, compulsion, etc. The notion of addiction is also applied to other human dependency-like characteristics and behaviors unrelated to drugs, such as attachment, gambling, sex and food consumption and other consummatory behaviors, possibly having overlapping neurophysiologic substrates. For instance, the central ghrelin signaling system, known as a regulator of eating behavior, is likely also involved in reward and associated dopaminergic activity related to various psychoactive substances (alcohol, cocaine, amphetamine) (Jerlhag et al., 2010). Impaired psychological self-regulation contributing to SUD risk may manifest in early-onset overeating, which in turn increases the rate of reproductive maturation (Must et al., 2005; Wang, 2002), resulting in a greater likelihood of affiliation with deviant/older peers and substance use. Leptin is involved in both fat metabolism and pubertal development in girls (Li et al., 2005), while its contribution to eating behavior is likely dependent on the cannabinoid system (Di Marzo et al., 2001). Numerous neurobiological mechanisms and respective genes expressed in the brain are potential contributors to variation in CLA, providing basis for testable hypotheses.

The genetic association studies so far have shown that the genes that may contribute to variation in the risk for addiction are not specific to a drug (e.g., rev. in Li and Burmeister, 2009). In some contrast to known or expected associations for the risk for disorders related to licit compounds (alcoholism and tobacco dependence; ADH and ALDH genes and CYP2B6 and CHRNA group genes), those related to illicit drugs point to dopamine, serotonin and GABA receptor and/or transporter and other non-drug-specific genes involved in neurotransmission and neurobiology (e.g., BDNF, ANKK1 and NRXN1). Even the connections of the nicotinic acetylcholine receptor genes, however, originally observed for nicotine dependence, have now been extended to cocaine and opiate dependence, sometimes with flip-flops of alleles specifically associated with the licit and illicit substances (Sherva et al., 2010).

Associations with CLA may be also expected for genes that are part of the mechanisms of neural plasticity, potentially involved in reaction to all external stimuli including (nonspecific) drug response. Such mechanisms likely include, for instance, epigenetic regulation, e.g., histone acetylation, which appears to participate in long-lasting neural adaptations related to stress and non-specific drug sensitization and possibly dependence (Covington et al., 2009; Romieu et al., 2008; Sanchis-Segura et al., 2009).

Panksepp et al. (2002) note the possibility that “narcotic addiction operates partially through mechanisms which ensured mammalian social bonding over the course of evolution” (p. 463). The relevance of the affiliative behaviors to the addiction risk is supported by the data related to the vasopressin–oxytocin system, including the arginine-vasopressin receptor 1A (AVP1A) gene. Studies in prairie and montane voles showed that interspecies differences in mating behavior (monogamous and polygamous, respectively) are determined by the brain distribution of this receptor (Insel et al., 1994). This receptor, as well as the vasopressin–oxytocin system in general, has been shown to be involved in attachment, affiliation and parental and reproductive behavior (rev. in Insel, 2010). This involvement demonstrates sex dimorphism, in that vasopressin (AVP) is particularly influential in males, while oxytocin (OXT) in females (although it seems unlikely that this sex-specificity is absolute). Consistent with this dimorphism, association of polymorphisms in the AVP1A gene with the risk for addiction has been observed in men but not women (Maher et al., 2011). Confirming the role of affiliative behaviors and pointing to a role of the biology of socialization in addiction, this association was mediated by indicators of the quality of mates’ relationship. Interestingly, a flip-flop of alleles of the associated polymorphism was observed for this gene as well, possibly explained by phenotypic differences in the samples related to the dual role (and respective differences in the severity) of drug use: as a prosocial facilitator and as an indicator of antisociality. The epigenetic process of DNA methylation may control the persistent behavioral effects of environmental factors mediated by the AVP receptor expression (Murgatroyd et al., 2009). The ability of the organism to retain such effects for a long time is shared with memory and learning. The mechanistic data, from the genetic level to higher nervous activity to complex behaviors, point at the connections of drug use behavior and addiction in the general framework of behavioral evolution.

3.4. Evolutionary roots of addiction

3.4.1. Rationale for evolutionary approach. As famously put by Dobzhansky (1973) in the title of his popular paper, “Nothing in biology makes sense except in the light of evolution”. In this section, we propose that, applied to addiction, the evolutionary outlook may not only shed light on its foundations and allow deeper insight into driving forces behind phenomena leading and attendant to drug abuse, but also puts them in a wider biopsychological perspective under the rubric of Darwinian, evolutionary medicine. Indeed, an evolutionary account of liability to addiction is necessary, because, on its surface, drug abuse/addiction presents an evolutionary paradox of voluntary self-inflicted fitness decay. As with other self-destructive or anti-reproductive behaviors (e.g., suicide, voluntary celibacy, homosexuality), plausible Darwinian explanations for this paradox need to be given for drug abuse and liability to addiction.

In fact, in contrast to the obviously negative influence of celibacy on fecundity and thus Darwinian fitness, it is not a priori certain that reproductive disadvantage is strong for drug use. Direct comparisons of fertility/fecundity between affected and nonaffected individuals may be faulty because of inability to ensure whether any differences observed are due to the factors acting pre- or post-initiation of substance use. In addition, drug use may be not so strongly related to fecundity as to be subjectively perceived as threatening it, and thus is even less under selective pressure than celibacy, which selection has not been able to preclude. Notably, inasmuch as pleasure in humans has evolved to become substantially detached from its likely original evolutionary goal of
signaling a fitness benefit and becoming a goal par excellence instead of being a means, pleasurable sensations, including those drug-induced, may override and erode what remains from that original connection and related perceptions. Risky sexual behavior and promiscuity associated with drug abuse also make it difficult to ignore a potential direct fitness benefit due to drug abuse that may obtain under certain conditions. This is consistent, for instance, with the association between parental addiction and faster sexual maturation, behavior dysregulation and addiction in children (Kirillova et al., 2001, 2008).

A cue for adaptive explanations was suggested by Darwin himself (Darwin, 1871), who, in a statement strikingly and diversely germane to the topic discussed herein, considered “liability to the same diseases . . . our tastes in common for the same stimulants, and the similar effects produced by them, as well as by various drugs” (vol. 1, p. 191) as a major evidence for common evolutionary origin of humans and other “higher mammals” (he fittingly listed taste for tea, coffee, “spirituous liquors” and tobacco smoking in monkeys [p. 12]). Lacking the ability to directly observe the process of evolutionary change, we can, following Darwin, rely on evolutionary stability, commonality of characteristics between humans and other species. Similarity of the structure of extremities between vertebrates or “our close similarity in minute structure and chemical composition” (Darwin, 1871, vol. 1, p. 191) with other mammals indicate both common origin and subsequent evolutionary divergence. Likewise, tracking behavioral characteristics, particularly those pertaining to addiction, to their common origin supports the evolutionary scheme.

The evolutionary approach allows addressing the problem in a reduced form. This reduction, in turn, enables drawing a complex human behavior to its more elemental psychological foundations, to the level of emotions, and, ultimately, to identifying their biological substrate. Addiction, therefore, when viewed from the perspective of behaviors involved, opens a window to a much wider area of human functioning than merely psychopharmacologic response to drugs including physiological dependence.

3.4.2. Common metric system. Evolutionary tracking of general mechanisms of addiction liability inevitably reaches the level of basic engines of behavior, emotions. Emotions, reflecting perceived reward and punishment, incentives and deterrents, are also the “common fitness metric across different stimuli . . . along which everything, from apples to oranges to cocaine, can be compared” (Panksepp et al., 2002, p. 460). It is illustrative that the brain areas responsible for extinction of fear, one of the primary emotions, and addiction overlap (Peters et al., 2009). The downside of having evolved such common fitness scale is that the brain has also become prone to evaluating cocaine to be as good or, indeed, much better than “apples and oranges”. Whereas the mechanisms underlying the CLA are grounded in this ancient and highly conserved system, which determines incentive motivation and reward-pursuing behavior, drugs produce a false signal of fitness benefit that is very potent (Nesse and Berridge, 1997). Drugs circumvent the complex modulators determining input into the system that assigns a fitness value to an action or experience, by acting upon the system itself. The input provided by drugs may thus be beyond the measurement system's usual range of stimuli, and has no relationship to what these inputs are supposed to represent.

This system may thus attach an extremely high value to the drug-related behaviors, forming foundation for psychological dependence. Similar deranging influences may be at work with refined high-calorie food products such as sugar or alcohol that takes a less defined position between nutrients and drugs. Their consumable quantities that are currently available far exceed those found in natural products that have historically supplied humans with these nutrient/energy sources (or, as it may be in the case of alcohol in addition to its psychoactive effect, protection of these sources from spoilage). Drug-related negative fitness signals, at least before full possible health consequences of drug use develop, are largely generated by the changeable societal norms, thus having to undergo cognitive processing modulated by numerous individual factors and microenvironment. Therefore, these negative signals only indirectly act upon the measurement system and may not be able to offset the powerful false-positive signal that substance consumption produces directly at the basic neurobiological level.

An important component of the neurobiological basis of this measurement system is likely the mesocorticolimbic dopaminergic circuitry. The dopaminergic system (DS) has been for a long time considered the “reward system.” It is also possible, however, that mesolimbic dopamine neurons code for incentive salience in general (Robinson and Berridge, 1993) or stimuli with high motivational impact pertaining to novelty, aversiveness, or deprivation (Bassareo and Di Chiara, 1997) and thus relevant to evaluation on the fitness scale. Among such stimuli, social interactions, attachment, parental behavior, regulated by the vasopressin/oxytocin release, may play an important part, connecting social behavior to addiction liability via common neurobiology (Insel, 2003), which is not necessarily limited to the DS.

Considering the ubiquity of addictive substances, the question may indeed arise as to why drug addiction, while relatively frequent, is not universal. The answer is perhaps the same as can be given to a similar question about overeating and subsequent obesity under the current conditions of virtually unlimited access to food in developed and even some third world countries. This access is a very recent environmental change. Obviously, mass access to abundant food is outside of what used to be the common food environment. The brain system controlling hunger and satiation that has evolved to motivate eating behavior determines the consumption limits, which, while varying in the population, generally determine the level of consumption that is compatible with or optimal for body function in what is termed the environment of evolutionary adaptedness, EEA (Bowlby, 1969). The evolutionary adaptation to food abundance, due to its recency, has not occurred to the degree that could prevent overconsumption and concomitant obesity and diabetes, even though they affect fertility and thus are under selective pressure. These conditions are well known to have reached epidemic proportions. Nevertheless, a substantial proportion of the population – whether via conscious effort or naturally reaching satiety commensurate with the actual energy and nutrient needs – is able to maintain a normal body mass.

Hedonic effects also have a ceiling, perhaps evolved due to resource limitations (Lende and Smith, 2002). This ceiling, high as it may be, is sufficiently low in many individuals not only to preclude continuous consumption (of food, alcohol, drugs, sex, etc.), but also to be outbalanced by negative effects to keep a consumatory behavior under control or even abolish it. Consumption of alcohol – a natural and ancient food component – is frequently limited by the noxious response that reduces the risk for incapacitation and addiction, assisted in humans by the societal norms that frequently restrict inebriated behavior (up to its full prohibition) and sometimes ritualize consumption of psychoactive substances (also up to its prohibition). Acute and developing tolerance further decreases hedonic effects, which may be an evolved, albeit no longer particularly effective (even acting in the opposite direction), mechanism to disincentivize further consumption of psychoactive substances. It is possible, however, that this mechanism contributes, along with the above discussed factors, to the not-so-seldom observed controlled use of alcohol and drugs as well as “maturing-out” of addiction, which occurs as it may with any other behavior when the perceived cost–benefit balance shifts to the cost perception due to age, career demands, family obligations, etc.

Subjective perception of cost (and benefit), of course, also varies individually, and the perception of “rock-bottom”, if any, ranges from a life-threatening health problem to a mere understanding that one’s volitional control of behavior has degraded. Feynman (1985), the Nobel prize-winning physicist, describes how he quit drinking upon getting frightened about experiencing “strong feeling that you have to have a drink” (p. 204). Similar aversive effects may be produced by an experience of the short-term memory lapse resulting from cannabis use, or by observation of personal/social degradation concomitant to drug use, when sufficient value is attached to the memory capacity or to the health or social status. Age- (and chronic drug use-) related changes may include the gradual lowering of the general hedonic effect and craving. These are testable hypotheses with potential practical implications. The common fitness metric/etiologic mechanisms may determine similarity between the patterns of development of habitual consummatory behaviors, their persistence, fluctuations and discontinuation, parallelisms between them, and comorbidities of their extreme forms (let alone comorbidities of substance-specific addictions) comprising spectra of disorders.

The question remains, why the hedonic ceiling that is achievable with drug use is so high? Whereas evolution has not dealt with the amounts of drugs consumed, neither could it have provisioned for the future use of morphine, heroin, vodka, cocaine or hashish to enable such a high amplitude (and thus a very wide range) of their brain effects. Why are there in place mechanisms that allow development of craving, be it drugs or food (positive and negative reinforcements are both involved)? The latter, as well as other mania-like cravings supported by respective hedonic states and unrelaxed to drugs (e.g., gambling and other high-risk activities, sexual, kleptomania), suggest the answer to these questions. These conditions are extreme variants of the behaviors that have evolved to support actions of direct importance to Darwinian fitness (reproduction, resource procurement), thus highly motivated and reinforced by pleasure. While high, this hedonic ceiling is within the norm of our species’ reaction to natural stimuli.

3.4.3. The amplitude of affective states (AAS) hypothesis. Moreover, as the hedonic states vary on a continuum from extreme unhappiness (dysphoria) to high euphoria, it is a positive affect change, rather than necessarily euphoria itself, that is generally sought. Indeed, the need to experience change has likely evolved to motivate the fitness-significant behavior. Humans are very seldom in a (pathological) state of constant hunger or other cravings. Both the basal state and the magnitude of this change are subject to individual variation, and satisfaction is frequently achieved by the changes resulting from behaviors within the socially acceptable/legal boundaries. Obviously, merely using an illegal drug (and alcohol before legal age) once or a few times is already a violation of legal boundaries, but frequently not quite the social ones, especially when only immediate environment, rather than the larger society, is concerned (family, peers, who may have more lenient or different views on substance use). Such use informs an individual about his personal physiological reaction, a component of motivation. Any psychoactive substance can serve this purpose, and it does depending on availability, opportunity, peer pressure, culture, etc. These episodes, and even regular “recreational” use, however, are per se usually insufficient to cross the threshold of gross violations of norms. In some cases, however, these boundaries are overstepped—whether because of a low individual basal affect/arousal level, a high individual hedonic ceiling, peer pressure or an otherwise caused need in a larger magnitude change than regularly achievable, particularly on the background of antisociality or otherwise lowering social boundaries. The amplitude of change that is sought is then above the safety level.

The following explanatory scheme may be proposed for the need to achieve such an amplitude. There is little doubt that the human nervous system has evolved under conditions where the danger was constant and even the nearest future uncertain. In particular, this concerns the elements of the environment that constitute resources—food, territory, reproductive or any other objects of consummatory behaviors. These danger and uncertainty, interspersed by the acts of consumption, determined a very wide range of affective states within the human adaptive norm, enabled by the effective mechanisms of coping with stressful conditions, by necessity—and according to their frequency—converting them into the baseline, the homeostatic level. It is historically only recently that relative stability and security have become possible for a large proportion of the population—largely in the developed countries. It stands to reason that such an improvement in the environment has generally resulted in an objectively diminished range, amplitude, of individually experienced affective states (AAS). The situation may be similar to the underloading of the immune system in the modern environment compared to the environment of evolutionary adaptedness, EEA (Bowley, 1969), as the potential cause of chronic inflammatory, autoimmune and allergic diseases (e.g., Rook, 2009). Analogous to the overactivity of the immune system in the absence of the agents to fight against in genetically highly predisposed individuals, some of them may seek ways to make up for lacking the wide fluctuations of stimulants (danger in particular) that the nervous system has evolved to experience. These individuals, which could be well adjusted in the EEA, are not optically adjusted in the current environment. Thus, their status of the nervous system is underarousal, resulting, e.g., in high novelty and sensation seeking (including that from substance use), risky and antisocial behavior, etc. This becomes emphasized particularly at transition to the reproductive period (which defines fitness) and relative independence, i.e., at adolescence. Cognition and related behavior control mechanisms have evolved and mature much later than the “basic instincts,” which include the need to consume large amounts of food fast when available, and store fat, potentially resulting in obesity.

Stressor underloading, resulting in a lower AAS, may call for the organism’s own compensation, directed either inward, as in the case of depression, especially bipolar disorder, or outward, as in risky activities. The latter naturally include behaviors restricted by the society, such as substance use and behavioral characteristics associated with the risk for addiction. This is not to say that antisocial or otherwise risky behavior has only recently appeared in the human repertoire. On the contrary—as mentioned in the case of food, what is currently viewed as a deviation could until recently have been a normative behavior with a clear instrumental value. The notion of deviation conflates a statistical term pertaining to the distance from the mean (or from a “normal” range of values, a tail of the distribution) and the attitude of the society that may or may not be based on objective and/or relevant criteria.

3.4.4. Drug abuse and (anti)social behavior. Adjustment to social norms is very important in humans, more than in other species because of advanced communication and procurement and defense of resources, which is strongly related to creating and maintaining social structures. Although acts of consumption are individual, participation in social groups (from family to state) is beneficial from the consumption standpoint inasmuch as it secures individual access to resources and protection. Shunning, marginalization, ostracism and exile are commonly perceived as extremely stressful in all social species. Conformity, compliance with the rules, let alone the ability to dictate the rules, allow joining (or, as in the latter case, reforming) the society, and are thus commonly pursued and valued even when no direct or consciously recognized benefit is derived. In fact, in ancient Greece, exile of
individuals viewed marginal for various reasons formed a human scapegoat ritual and was considered a therapeutic, purifying measure for the society, perhaps harking back to the times when only “one member [of the pack], preferably a marginal, weak, or sick member, falls victim to the beasts [could] the others escape’ (Burkett, 2004, p. 84). Interestingly, the word that was used for the scapegoat, pharmakos (πομάκος), while etymologically related to magic (Harrison, 1903), is of the same root as, and perhaps a form of, pharmakon, drug or poison (Hughes, 1991), meaningfully echoing the hypothesized reasons for illicit drug use. With the notion of the outcast, castaway, embedded in human mentality, the societal rules are obeyed not only because of the conscious fear of punishment, but because they constitute part of the social structure and are thus assimilated indeed akin to Kant’s categorical imperative. (To be sure, this assimilation is imperfect, as humans are not a eusocial species like communal bees, with a biologically hard-wired social structure.) Similarly, factors other than lack of fear contribute to violations of those rules, and the covariation between such violations across early childhood evaluations is almost entirely of genetic origin (Petitclerc et al., 2011).

Marginalization, which turns a member of the society into a non-member, an Orwelian unperson, an alien, ultimately a non-human, can be triggered by any perceived difference from the mainstream society, let alone by the individual’s attacking the society. As illustrated by H.G. Wells in The Country of the Blind, even an objective advantage (of vision over blindness), instead of conferring the proverbial kingship, may become a handicap, as “margin” corresponds to anything with a low(er) frequency in the population (cf. the minority status in many societies). Marginalization, while sometimes religiously/culturally conditioned to call for compassion, may evoke the mainstream’s reactions ranging from indifference and passive cruelty, as recently illustrated by 25 people passing by a dying homeless person bleeding to death, some taking pictures (Livingston et al., 2010), to societally licensed and encouraged violence, mass murder and genocide. Although compassion, the “instinct for sympathy”, in Darwin’s (1871) view, could be checked only “for a contingent benefit, with an overwhelming present evil”, this instinct’s application may be circumscribed to “weak and helpless” (p. 136) and is heavily contingent on culture (according to what was then the zeitgeist, Darwin juxtaposed “savages” and “civilized men”).

As in ancient societies, marginalization in our days, whether real or perceived, leaves few options, which, if based on behavior disorder and problem behavior, are frequently related to aggression directed at self (auto-aggression, including suicide) or the society. The latter variety includes the behaviors of gang organization, whereby a surrogate form of society is created, enabling access to resources and group support (Vanyukov, 2004). Moreover, this quasi-society frequently builds on the structure and relations in the society’s tightest unit, the family—both literally (some gangs involve actual multigenerational family participation (Ruble and Turner, 2000)) and figuratively (gangs are often described by their members as families and operate as family systems, with expressions of loyalty and devotion usually reserved for close relatives). While antisocial vis-à-vis the society at large, the behaviors endorsed are prosocial as pertains to the group. Group actions, particularly for resource procurement or defense, bearing the stamp of group approval, are viewed as legitimate (quasi-legal). At the same time, it is exactly the behavior deviation characteristics that led to marginalization in the first place, related to behavior (dys)regulation, that become normative and promoted in the gang environment as a constructive necessity: because the goal is to recreate a society, the differences from the mainstream tend to be viewed with the sign opposite to the mainstream’s. Notably, behaviors characterized as rule-breaking (obviously, the mainstream society’s rules are meant) and aggressive are highly genetically correlated in children (Bartels et al., 2003).

Antisocial behavior converges in the same construct with drug dependence from adolescence to adulthood, and the heritability of this construct, at least in males, grows with age (Hicks et al., 2007). Behavior dysregulation/disinhibition and deviance become the core phenotypes in the quasi-society/family of a deviant group, and the ability to express them is the condition for membership (e.g., gang initiation rituals involving criminal acts). As noted by Mealey (1995), competitively disadvantaged youth may be seeking a social environment in which they may be less handicapped or even become superior. This is a potential source of homophily – phenotypic similarity of peers due to tendency for the phenotypically like non-relatives to aggregate in groups – that may to a large degree be based on direct behavioral phenotypic assortment, akin to phenotypic assortative mating. This may translate into an advantage for antisocial individuals and thus into their higher perceived Darwinian fitness—in deviant social groups (Vanyukov et al., 2003b). Inasmuch as antisociality is associated with substance abuse, the latter also turns into an indicator of fitness benefit. The existence of such groups is consequently perpetuated as long as they serve a means to convert what is considered a disadvantage in the contemporary mainstream society (and/or discouraged there) into a benefit. This allows the individual to take comfort in group protection, cooperation and reciprocal altruism, thereby increasing their perceived fitness (Vanyukov, 2004). While homophily may be one of the mechanisms facilitating prosocial behavior toward non-relatives, drug use may be one of the readily observable phenotypic indicators of the latent traits that such groups assert on, facilitating such assortment.

The effect of parental SUD liability phenotype on the child’s peers’ deviance, growing over time (due to both contemporaneous and lagged effects), is consistent with both growing disconnection from parents (which is likely to be contributed by the higher level of maltreatment from drug-using parents) and the increasingly active homophilic choice of deviant peers (Kirillova et al., 2008). Inasmuch as homophily results from the individual’s selection into, and/or his choice of, a behaviorally similar peer group, its genetic effects are similar to the effects of phenotypic assortative mating (Guo, 2006; Vanyukov, 2004). The phenotypic correlation between peers can induce higher genetic (in proportion to heritability) and environmental similarity between the unrelated members of the group, such as a gang. This would further augment the active genotype–environment correlation (Scarr and McCartney, 1983), insofar as the choice of environment (peer behavior) will be in part genetically mediated. Indeed, the choice of delinquent peers is significantly heritable (Button et al., 2007; Walden et al., 2004) and genetically correlated with conduct problems, reflecting active genotype–environment correlation compounded with the reciprocal influence of peers on the individual’s behavior, in effect rendering peers part of extended phenotype. It is conceivable that, considering human communication capabilities, the homophilic groups can also be virtual—not necessarily in the cyberspace or electronic media only, but via the perception of what behavior is appropriate given a certain degree and modality of one’s behavioral deviation. It is thus not necessary for an individual with that behavioral deviation (e.g., CD) to formally belong to a deviant group (e.g., a gang) in order to manifest the respective set of symptoms including drug use, in part guided by such an indirect peer phenotypic assortment process.

Whereas some detachment from parents is common in adolescence as a concomitant or part of the autonomization process, it is not per se directional but rather reflects passive disconnection that attains a directional momentum through both geno-/phenotype–environment correlation and the independent effect of environmental exposure. These formative influences may
act synergistically. As the (adolescent) behavioral options are relatively limited, the most available or individually valuable behavioral model is the one that is likely to be accepted. It has been observed that the severity and temporal stability of antisociality are related to whether its onset precedes puberty or coincides with it (Moffitt, 1993). Based on such timing variation, two categories of antisocial phenotypes were proposed, life-course-persistent and adolescence-limited. These categories take into account that an early onset of antisociality is frequently associated with its long-term character, whereas its onset in adolescence is almost normative and less frequently predicts its chronicity. As Moffitt (1993) reviews, “life-course–persistent persons miss out on opportunities to acquire and practice prosocial alternatives at each stage of development” (p. 683). Nevertheless, whereas a small proportion of cases may indeed have expressed “antisocial” behavior from infancy on, the onset ages likely cover the whole period preceding adolescence. This is, in fact, illustrated in the discussed paper’s Fig. 3, depicting a continuous age distribution rather than a typological dichotomy.

Early difficulties in assimilating social norms may be due to various reasons, both organicism and environmental. These groups of reasons are not independent of each other. O’Connor et al. (1998) found that 7–12-year old adopted youth at high risk for antisocial disorder (based on the biological parent status) evoked negative and coercive parenting from their adoptive non-antisocial parents. In turn, parenting style, as perceived by the child, is predictive of childhood disruptive behavior disorders (conduct disorder and attention deficit hyperactivity disorder) (Vanyukov et al., 2007), known to be frequent precursors of SUD. Importantly, for CD, the more potent predictor of SUD, this relationship has different directions for fathers and mothers. These relationships may also be influenced by a functional polymorphism in the MAOA gene coding monoamine oxidase A, a key enzyme in the metabolism of amine neurotransmitters (e.g., dopamine, serotonin). Although the individuals with early-onset antisociality/underarousal/low-AAS are a minority, by negating societal (parental) norms they constitute a highly charged social nucleus attracting adolescents who come into contact with it, real or virtual—unless protected genetically, by upbringing, and/or by social environment. Premature autonomy increases the risk of problem behavior (and thus of addiction) regardless of whether it can be ascribed to parental management deficits (Dishion et al., 2004) or relatively precocious physical maturation (combined with suboptimal brain maturation) (Tarter et al., 1999). Both are likely to result in affiliation with older and deviant peers, mediating transmissibility of addiction liability (Kirillova et al., 2008). The detachment from parents, augmented by the age-related rise in explorative behavior and expanded access to novel environments and stimuli, may include dissociation from the behavioral norms of the parental generation, i.e., frequently, the social norms.

Akin to how any perceived deviation may “serve” to turn an individual into a pharmakos, an outcast, any behavioral deviation may in principle serve as a basis of elevated liability to addiction. A relatively small deviation in childhood or adolescence, due to passive, evocative and then active phenotype (and genotype)–environment correlations (Scarff and McCartney, 1983) compounding its effect, may snowball to full-blown antisociality, with drug abuse and subsequent addiction as its attributes. It may be expressed in the need to self-medicate against depression, corresponding to the internalizing deviation, or to correct for an externalizing deviation, including a subjectively or objectively low amplitude of sensations, boredom, underarousal, wherein lies a possible connection with psychopathy (overlapping with antisociality). More than 50% of addicts have other psychopathology (Regier et al., 1990), and they are at particularly high risk of having comorbid externalizing disorder (Compton et al., 2005; Kessler et al., 2001; Warner et al., 1995). In genetically informative data, SUD clusters with externalizing rather than internalizing behavior disorders (Kendler et al., 2007). This pertains to genetic connections (correlations) as well, and gives grounds to consider addiction an externalizing spectrum disorder (Iacono et al., 2008; Krueger et al., 2002). In other words, mechanisms of variation in CLA substantially overlap with those determining variation in social behavior. In effect, drug use behavior and its consequences comprise a facet and an indicator of social behavior, frequently engendered and sustained by the latter’s earlier manifesting deviating forms, behavioral “gateways” to addiction.

To be sure, the externalizing–internalizing classification can be reduced to more elemental traits. This classification also does not entirely cover the domains of phenotypic and mechanistic variation contributing to individual differences in addiction liability. For instance, the mechanisms of attachment and bonding — an important component of socialization — potentially play an important role in addiction liability (Panksepp, 1998) and overlap at the neurobiological and genetic level (Insel and Young, 2001; Maher et al., 2011). Phenotypically, they share in common characteristics of dependence. It is hardly a coincidence that drug use is usually initiated in adolescence, when affiliation is relocated from parents to peers.

Germaine to addiction liability, this relocation is assisted by the relationship between parental SUD and faster pubertal maturation of their sons, and, on the other hand, between faster physiologic maturation and affiliation with deviant peers, novelty seeking, and development of conduct disorder and SUD (Kirillova et al., 2001, 2008). Both the testosterone level in boys and their affiliation with deviant peers at an earlier time point are cross-predictive of these variables’ values at a later time point, biasing development to conduct disorder and SUD, and in part mediating transmission of SUD risk from parents to children. These findings support a role of deviations in physiological maturation (perhaps mistimed with neural maturation) in the addiction risk (Tarter et al., 1998). Due to the association between earlier maturation and peer deviancy, social groupings of deviant peers (gangs) may be enriched with physiologically earlier maturing individuals, which is compounded with the peer pressure to display early “adult” behavior, including smoking, drinking and drug use. This indicates an important extension of the CLA phenotype, which may be tested and taken into account in prevention and intervention.

Because of the behavioral phenotypic and likely increased genetic intragroup similarity, the concept of inclusive fitness (Hamilton, 1964) is applicable to the gang situation. Inclusive fitness involves not only the reproductive success of the individual, but also that of relatives who share his genes. The concept of inclusive fitness (as well as group selection and evolutionary explanation of altruism) originates from Darwin (1871), who noted that [a] tribe including many members who, from possessing in a high degree the spirit of patriotism, fidelity, obedience, courage, and sympathy, were always ready to aid one another, and to sacrifice themselves for the common good, would be victorious over most other tribes; and this would be natural selection (Part 1, p. 166).

As noted above, genetic similarity between gang members with similar phenotypes may indeed be higher than that among the non-associated individuals from the general population. The plausibility of extension of the notion of kin to non-relatives is illustrated by the well-publicized effective sexual taboo not only among close relatives (incest taboo), but, independent of sibship, among children raised together, as in kibbutzim (Sheper, 1971; Wilson, 2000). The genetic correlations due to homophily may thus be the ultimate evolutionarily significant goal of such group associations, because it increases the probability of survival and reproduction of identical genes—without the negative consequences of inbreeding where genetic correlations are primary, rather than secondary as under assortative mating or homophily. The genotype-driven search for permissive environment, active genotype–environment
correlations, could be a means to achieving this goal. Therefore, regardless of whether there are blood relatives in the group, altruistic behavior toward the gang members may be consistent with the induced genetic correlations between non-relatives and the extension of inclusive fitness to them. It is noteworthy that this group behavior has been shown to depend on the oxytocin–vasopressin system. In particular, administration of oxytocin in males (only males were studied) elevated in-group cooperation and defensive aggression (De Dreu et al., 2010). It seems likely that other components of the stress-socialization system subserved by vasopressin, oxytocin and stress-reactivity mechanisms participate in both the formation and individual variation of social behavior.

The fact that illicit drug abuse is part of the repertoire of problem/antisocial behavior rooted in human behavioral evolution, along with the other commonalities discussed, directs attention to the area where interventions may be particularly effective. The distinction between the sources of variation – drug-specific or common – is obviously important for elucidation of etiologic factors, including genes. An important conclusion from the substantial non-specificity of genetic mechanisms of phenotypic variation in addiction liability is the likely predominant significance of behavior regulation, particularly as it pertains to social behavior, and nonspecific neurobiological pathways, as opposed to the specific mechanisms of drug biotransformation or action. Although such specific mechanisms do exist and, moreover, can be capitalized upon in intervention (the therapeutic effects may be large even if the natural effects of the same factors are small), their role appears to be minimal under “natural” conditions. Therefore, there may be a greater chance to discover actionable mechanisms when nonspecific liability, shared in common between risks for addiction to specific drugs, is addressed. Such interventions, channeling ontogenesis of the CLA phenotype away from the threshold, would target the entire range of its manifestations, including those unrelated to drugs directly. Whereas the “gateway hypothesis” assumes the progression related to, if not caused by, a “gateway” drug, with the consequent prevention and policy implications targeting drug use as such, the CLA theory focuses on the etiologic processes at all levels of biological organization that result in drug abuse and addictive behavior.

4. Conclusions

To summarize the discussion of the two competing concepts, “gateway” sequence and common addiction liability, the following points can be reiterated. Juxtaposed with the parsimonious and empirically proven concept of CLA, explaining comorbidity of substance use disorders and polydrug abuse by commonality in etiologic mechanisms, the GH appears redundant. Considering the numerous exclusions from and the triviality of the “typical” substance use initiation sequence, as well as its irrelevance to substance use disorders, its utility is uncertain. Taking into account the danger that the presentation of that sequence as a “theory” may be (and has been) erroneously interpreted in causal terms, it may be hindering both research and intervention. In contrast, CLA, a behavioral/psychological trait, manifests in a range of “gateway” behaviors grounded in the mechanisms of socialization and affective/cognitive regulation with deep evolutionary roots. The CLA concept provides theoretical and empirical foundation to research in etiology, quantitative risk and severity measurement, and targeted non-drug-specific intervention. This conclusion is supported by the papers in this special issue.

Role of funding source

NIDA had no role in the analysis, preparation, and writing of the report, or in the decision to submit the paper for publication.

Contributors

All authors have materially participated in the manuscript preparation. Michael Vanyukov wrote the first draft of the manuscript. All authors contributed to and have approved the final manuscript.

Conflict of interest

All authors declare that they have no conflicts of interest.

Disclaimer

The views and opinions expressed in this report are those of the authors and should not be construed to represent the views of NIDA or any of the sponsoring organizations, agencies, or the U.S. government.

References


